Best Available Copy

ADA130702

COMPONENT PART NOTICE

THIS PAPER IS A COMPONENT PART OF T' ; FOLLOWING COMPILATION REPORT:

(TITLE): Proceedings of the Annual Confe ..ce of the Military Testing Association (23rd) held at Arlington, Virginia 25-30 October 1981. Volume 1.

(SOURCE): Army Research Inst. for the Behavioral and Social Sciences, Alexandria, VA

TO ORDER THE COMPLETE COMPILATION REPORT USE AD-A130 702.

THE COMPONENT PART IS PROVIDED HERE TO ALLOW USERS ACCESS TO INDIVIDUALLY AUTHORED SECTIONS OF PROCEEDINGS, ANNALS, SYMPOSIA, ETC. HOWEVER, THE COMPONENT SHOULD BE CONSIDERED WITHIN THE CONTEXT OF THE OVERALL COMPILATION REPORT AND NOT AS A STAND-ALONE TECHNICAL REPORT.

THE FOLLOWING COMPONENT PART NUMBERS COMPRISE THE COMPILATION REPORT:

COMPONENT PART NOTICE (CON'T)

AD#:	TITLE:
AD-P001 304	Computerized Adaptive Testing: From Psychometrics to Systems.
AD-P001 305	Development of a Work Behavior Survey for Thirty-Five Federal Clerical Occupations.
AD-P001 306	CODAP80 (Comprehensive Occupational Data Analysis Programs): The New Occupational Analysis Computer System.
AD-P001 307	Current Directions in Selection Research for the Royal Navy.
AD-P001 308	An Evaluation of the Fairness of the Flight Aptitude Selection Test (FAST).
AD-P001 309	The 'New' Army Officer Evaluation Reporting System.
AD-P001 310	NROTC (Naval Reserve Officers' Training Corps) Joins the BIG Test Movement.
AD-P001 311	Work Characteristics: A Task-Based, Benchmark Approach.
AD-P001 312	Interactive Video for Interpersonal Skills Training.
AD-P001 313	Testing Recall by Means of Multiple Choice Tests.
AD-P001 314	Performance Criteria Development for Army Field Recruiters.
AD-P001 315	The Effects of Job Satisfaction on Air Force Enlistee Retention.
AD-P001 316	Differential Effectiveness and Efficiency of Individualized Instruction. II Major Findings.
AD-P001 317	An Enriched User Oriented Cost and Training Effectiveness Analysis Methodology.
AD-P001 318	Prediction of Boot Camp Attrition: IRT versus Number-Right Scoring.
AD-P001 319	Soldier/Machine Interface Evaluation of Army Developmental Systems.
AD-P001 320	Academic Preparation, Aptitudes Measures and Army Officer Performance.
AD-P001 321	Army Officer Success in the Different Types of Career Branches.
AD-P001 322	Use of CODAP (Comprehensive Occupational Data Analysis Programs) System 80 Task Modules in Occupational Analysis.
AD-P001 323	A Monte Carlo Comparison of Three Optimal Test Scoring Procedures.
AD-P001 324	Air Force Civilian Promotion Appraisal System Development.
AD-P001 325	Exploration of Instructional Application of Videodisc Technology in the Marine Corps.
AD-P001 326	Testing during Training: Why does it Enhance Movement Retention.
AD-P001 327	Differential Effectiveness Efficiency of Individualized Instruction. 1. Background and Study Design.
AD-P001 328	Relating Selection Predictions to Attrition in the British Infantry.
AD-P001 329	Counter-Attrition Programs in the US Military.
AD-P001 330	Trend Analysis and Projections of the Population and Race/Sex Composition of the Federal Criminal Investigating Series.

COMPONENT PART NOTICE (CONT'D)

AD#:	TITLE:
AD-P001 331	The Strategic Weapon Training Program. Part 2. Executive Steering Group's Role. SWS Personnel and Training Evaluation Program.
AD-P001 332	Effect of Expectations on Attrition in Canadian Forces Recruit Training.
AD-P001 333	The Honor Code: The Optimum Criterion of Honor.
AD-P001 334	Improving Acceptance of Automated Counseling Procedures.
AD-P001 335	Enlistment Screening Test to Predict Army Aptitude Composite Scores.
AD-P001 336	The Great Training Robbery.
AD-P001 337	The Use of the Armed Services Vocational Aptitude Battery (ASVAB) as a Predictor of Early Discharge from the U.S. Army.
AD-P001 338	Applying Linguistic Theory to Job-Related Verbal Tests.
AD-P001 339	Commanders' Assessment of Unit Effectiveness Measures.
AD-P001 340	The Measurement of Morale.
AD-P001 341	Early Identification of Academic Difficulties in Foreign Language Training.
AD-P001 342	Computer-Based Skill Qualification Testing: Development and Validation.
AD-P001 343	An Investigation into the Use of an Aircraft Landing Test (ALT) in Aircrew Selection.
AD-P001 344	Flying Selection in the Royal Air Force.
AD-P001 345	Gender Differences in the Aircraft Maintenance Career Field.
AD-P001 346	Utility of the Field Dependence Dimension in Military Settings.
AD-P001 347	T & E (Training and Experience) Rating with Objectively Scored Applicant Appraisal Questionnaire.
AD-P001 348	An Attempt to Validate a Short Job Evaluation Questionnaire.
AD-P001 349	Reducing Adverse Impact via a Measure of Applicant Disadvantagedness.
AD-P001 350	Development and Calibration of Enlistment Screening Test (EST) Forms 81a and 81b.
AD-P001 351	Twenty Years Later: A Follow-up of the Officer Evaluation Center.
AD-P001 352	The Problem of Range Restriction in Test Validation.
AD-P001 353	Determinants of Readiness for Reenlistment in the Federal Armed Forces.
AD-P001 354	Task Analysis for Weapons Systems Testers: Shortcut to Paydirt in Inflationary Times.
AD-P001 355	An Occupational Analysis of Air Force Behavioral Scientists: Preliminary Results.
AD-P001 356	Performance and Skill Requirements for Fire Support Teams.

PROCEEDINGS

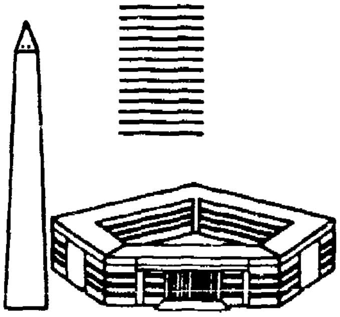
23rd Annual Conference
of the

MILITARY TESTING ASSOCIATION

Volume I

Coordinated by the

U.S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES



ARLINGTON, VIRGINIA 25-30 OCTOBER 1981

PROCEEDINGS

23RD ANNUAL CONFERENCE

of the

MILITARY TESTING ASSOCIATION

coordinated by

US ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

ALEXANDRIA, VIRGINIA

VOLUME I

Pentagon City - Quality Inn

Arlington, Virginia

25 - 30 October 1981

TABLE OF CONTENTS

1981 MTA Conference Staff

		Assisted by:
President	COL L. Neale Cosby	
Conference Coordinator	Arthur Marcus	
Secretary	Joyce L. Shields	
Program Committee	M. A. Fischl	Donna C. Angle
	Arthur C. F. Gilbert	Francis E. O'Mara
		Laurel Oliver
		Michael G. Rumsey
		Nancy Tulloh
Membership/Registration Committee	Angelo Mirabella	Virginia Rua
	Robert M. Ross	Anna J. Sapp
Finance Committee	MAJ David C. Kirk	
Graphics Aids/Publications Committee	Judah Katznelson	Joyce Ardale
		Ben Baggett
		Bonny Creskoff
		David S. Hoffman
		Marcia Mazur
		John H. Nagel
Social/Logistics Committee	Paul A. Gade	
	Jack M. Hicks	
Public Relations Committee	Robert B. Iadeluca	

PROGRAM

23rd ANNUAL CONFERENCE OF THE MILITARY TESTING ASSOCIATION

Arlington, VA

MONDAY, 26 OCTOBER 1981

PRELIMINARY ACTIVITIES

0900 Steering Committee Meeting
(Parlor B)

1030 Visitor Orientation
(Capital View Ballroom)
"The Washington Brief": Information on sightseeing, restaurants, and shopping, as well as current theater reviews, will be presented for the benefit of all registrants.

1330 CONFERENCE OPENING AND PLENARY SESSION
(Capital View Ballroom)

Opening Remarks

COL L. Neale Cosby, Commander, US Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA

Welcome

Joseph Zeidner, Technical Director, US Army Research Institute for the Behavioral and Social Sciences, and Chief Psychologist of the Army

Keynote Address

"Manning the Battlefield of the 1980s and 1990s." LTC Paul F. Gorman, Assistant to the Chairman, Joint Chiefs of Staff, Washington, D.C.

Special MTA Project Report

"Writing and Publishing the Book." Raymond O. Waldkoetter, US Army Research Institute for the Behavioral and Social Sciences, Fort Sill, OK

MTA Business Office: Parlor A
MTA Hospitality Suite: Penthouse 1401; Hours: 1600-2100 Sun., 1700-2200 Mon.-Wed.

TUESDAY, 27 OCTOBER 1981

	SELECTION AND PSYCHOMETRICS (Capitol View Ballroom East)	OCCUPATIONAL RESEARCH (Capitol View Ballroom West)	TRAINING AND PROFICIENCY MEASUREMENT (Terrace I)	ORGANIZATIONAL FACTORS (Terrace II)
0830	Measurement Methodology I Panel. Profiling the Aptitudes of the Current Youth Population Chair: Wayne S. Sellman, Office of the Assistant Secretary of Defense/MRA&L, Washington, DC Mark J. Eitelberg, Janice Laurence, Brian K. Waters, HumRRO, Alexandria, VA	Task Analysis I Paper Session; Chair: John L. Miles, Jr., ARI The Job/Task Analysis/Skills and Knowledge Marriage. Thomas H. Ansbro, William A. Hayes, Office of the Chief of Naval Education and Training, Pensacola, FL Task Analysis for Weapons System Testers: Shortcut to Paydirt in Inflationary Times. John L. Miles, Jr., US Army Research Institute, Alexandria, VA Simulator Instructor Task Analysis for Console Design. MAJ Robert P. Bateman, USAF (Ret.), Systems Research Laboratories, Inc., Dayton, OH The Identification of Common Duties in 13 Federal Technical Occupations. Stephen M. Cormier, Amiel T. Sharon, Cynthia Diane, Office of Personnel Management, Washington, DC	Training Issues I Paper Session; Chair: Melissa S. Berkowitz, ARI A New Direct Support Maintenance System. Douglas I. Bobko, US Army Research Institute, Alexandria, VA CHAPARRAL Training Effectiveness Analysis. John D. Tubbs, US Army TRADOC Systems Analysis Activity, White Sands Missile Range, NM Differential Effectiveness and Efficiency of Individualized Instruction: I. Background and Study Design. Eugene R. Hall, Jon S. Freda, US Navy Training Analysis and Evaluation Group, Orlando, FL Differential Effectiveness and Efficiency of Individualized Instruction: II. Major Findings. Jon S. Freda, Eugene R. Hall, Larry H. Ford, US Navy Training Analysis and Evaluation Group, Orlando, FL	Women in the Military I Panel: Women in the Military Chair: Nora S. Kinzer, Industrial College of the Armed Forces, Washington, DC LT D. G. Forescell, Canadian Forces Personnel Applied Research Unit, Willowdale, Ontario; David R. Segal, University of Maryland, College Park, MD; Mady W. Segal, Walter Reed Army Institute of Research, Washington, DC; CPT Susanne F. Simpson, Colts Neck, NJ
1000	Coffee Break	Coffee Break	Coffee Break	Coffee Break
1030	Measurement Methodology II Paper Session; Chair: Arthur C. F. Gilbert, ARI Reducing Adverse Impact through a Measure of Applicant Disadvantagedness. Walter C. Mann, Jr., Office of Personnel Management, Washington, DC Testing Recall by Means of Multiple Choice Tests. Ralph Rhwall, Command and General Staff College, Ft. Leavenworth, KS	Task Analysis II Paper Session; Chair: R. M. Chambers, ARI Performance and Skill Level Requirements for Fire Support Teams. John A. Modrick, T. A. Plocher, J. D. Hutcheson, Honeywell, Inc., Minneapolis, MN; R. M. Chambers, US Army Research Institute, Alexandria, VA Human Aptitude/Ability Assessment Techniques for System Designers. Paul G. Rossmeissl, Stanley J. Kostyla, James D. Baker, US Army Research Institute, Alexandria, VA	Performance Appraisal Paper Session; Chair: Jack H. Hicks, ARI A Performance Management System for the USCG: New Way of Doing Business? CDR Nicholas H. Allen, Bradford F. Sharp, US Coast Guard Headquarters, Washington, DC A Performance Management System for the USCG: Strategy for Development and Implementation. Bradford F. Sharp, CDR Nicholas H. Allen, US Coast Guard Headquarters, Washington, DC The Air Force Job Performance Appraisal System. Nancy A. Thompson, Douglas K. Cowan, John A. Guerrieri, Air Force Human Resources Laboratory, Brooks AFB, TX	Attrition and Retention I Paper Session; Chair: Glenda Y. Nogami, ARI Forecasting Army Officer Retention Prior to Commissioning. Claude F. Bridges, John W. Houston, Richard P. Butler, US Military Academy, West Point, NY Determinants of Readiness for Reenlistment in the Federal Armed Forces. Klaus J. Puzicha, Adelheid B. Meissner, German Armed Forces Psychological Services Research Institute, Bonn, Germany
1130	The Navy Personnel Accessioning System. William A. Sands, Navy Personnel R&D Center, San Diego, CA A New Approach in Developing Test Scales. Claude F. Bridges, US Military Academy, West Point, NY	Development of a Work Behavior Survey for 35 Federal Clerical Occupations. Cynthia C. Diane, Amiel Sharon, Stephen Cormier, Office of Personnel Management, Washington, DC Identifying Common Duties Among Naval Skilled Trades. Amiel T. Sharon, Cynthia Diane, Stephen Cormier, Office of Personnel Management, ...	Analyzing the Performance Rating Process of Expert and Nonexpert Raters. Cristina C. Banks, University of Texas, Austin, TX	Panel: Decision Aids for Personnel Actions. Chair: Joe H. Ward, Jr., Air Force Human Resources Laboratory, Brooks AFB, TX Neil B. Dumas, US Army Research Institute, Alexandria, VA; ...

Copy available to DTIC does not permit fully legible reproduction

1200	Lunch	Lunch	Lunch	Lunch
	Measurement Methodology III	Occupational Requirements	Training Issues I	Attrition and Retention II
1330	Panel. Psychometric Considerations for Adaptive Testing Systems: Test Development and Calibration	Paper Session; Chair: Stanley F. Bolin, ARI	Paper Session; Chair: J. D. Hagman, ARI	Paper Session; Chair: Guy L. Siebold, ARI
	Chair: Isaac I. Bejar, Educational Testing Service, Princeton, NJ	Utility of the Field Dependence Dimension in Military Settings. Robert Loo, Canadian Forces Personnel Applied Research Unit, Willowdale, Ontario, Canada	Utilization and Training Workshops: A USAF Innovation in the ISD Process. MAJ Frank C. Gentner, LTC Jimmy L. Mitchell, USAF Occupational Measurement Center, Randolph AFB, TX	Effects of Expectations on Attrition in Canadian Forces Recruit Training. CPT D. S. Norton, Canadian Forces Personnel Applied Research Unit, Willowdale, Ontario, Canada
	Neil Dorans, Carol Dwyer, James Douglass, Educational Testing Service, Princeton, NJ; Thomas Warm, US Coast Guard Institute, Oklahoma City, OK	Forecasting Federal Personnel Requirements Through Trend Analysis. Hanoch Rothstein Hirsh, Office of Personnel Management, Washington, DC	Testing during Training: Why Does It Enhance Movement Retention? Joseph D. Hagman, US Army Research Institute, Alexandria, VA	An Empirical Investigation of First Term Attrition. Glenda Y. Nogami, US Army Research Institute, Alexandria, VA
		An Enriched User Oriented Cost and Training Effectiveness Analysis Methodology. Edward W. Frederickson, Applied Science Associates, Inc., El Paso, TX	The Great Training Robbery. Judah Katznelson, US Army Research Institute, Alexandria, VA	Attrition: Causality, Explanation, and Level of Analysis. Guy L. Siebold, US Army Research Institute, Alexandria, VA
		The Development and Application of Measures of Occupation Learning Difficulty. Joseph L. Weeks, Air Force Human Resources Laboratory, Brooks AFB, TX	Designing and Managing Technical Training Programs: A Learner Based Model. Brandon B. Smith, University of Minnesota, Minneapolis, MN	Counter-Attrition Programs in the US Military. Jack M. Hicks, US Army Research Institute, Alexandria, VA
1530			Coffee Break	Coffee Break
1545			Training Issues II	Women in the Military II
			Paper Session; Chair: Joan Harman, ARI	Paper Session; Chair: Jack M. Hicks, ARI
			Training and Cost Effectiveness of Two Avionics Maintenance Training Simulators. Louis F. Cicchinelli, and Ken R. Harmon, Denver Research Institute, Denver, CO	Women: An Underutilized Resource. Mark Thiessen, Operational Research and Analysis Establishment, Ottawa, Ontario, Canada
			Models of Human Information Processing: Implications for Trainer/Simulator Design. Spencer C. Thomason, Essex Corporation, White Sands Missile Range, NM	Gender Differences in the Aircraft Maintenance Field. M. Suzanne Lipscomb, Air Force Human Resources Laboratory Brooks AFB, TX
				Performance of Male and Female Cadets during Cadet Field Training. MAJ Jerome Adams, US Military Academy, West Point, NY; Jack Hicks, US Army Research Institute, Alexandria, VA

1800	Optional evening tour of Washington (Two-plus hours) with champagne and hors d'oeuvres. Scenicruiser bus pick-up and return to hotel or drop-off in historic Georgetown, known for its fine restaurants...$15 per person

WEDNESDAY, 28 OCTOBER 1981

	SELECTION AND PSYCHOMETRICS (Capital View Ballroom East)	OCCUPATIONAL RESEARCH (Capital View Ballroom West)	TRAINING AND PROFICIENCY MEASUREMENT (Terrace I)	ORGANIZATIONAL FACTORS (Terrace II)
0800	Adaptive Testing Panel: Psychometric Considerations for Adaptive Testing Systems: Administration and Validation Chair: Isaac I. Bejar, Educational Testing Service, Princeton, NJ Douglas H. Jones, Donald A. Rock, Howard Wainer, Educational Testing Service, Princeton, NJ, Arlen Waters, HumRRO, Alexandria, VA	CODAP I Paper Session; Chair: Raymond O. Waldkoetter, ARI CODAP - An Application to Training Information System Analysis. LTC Derek Beel (Ret.), Royal Naval School of Educational and Training Technology, Portsmouth, England Use of CODAP System 80 Task Modules in Occupational Analysis. Doug Goodgame, Texas A&M University, College Station, TX CODAP: Analytical Tool for Position Classification? Jean Newton, Office of Personnel Management, Washington, DC CODAP 80: The New Occupational Analysis Computer System. Richard W. Dickinson, Texas A&M University, College Station, TX	Performance Appraisal II Paper Session; Chair: Timothy W. Elig, ARI Training Evaluation - A Pragmatic Overview. Gilbert P. Yard, Royal Canadian Mounted Police, Ottawa, Ontario, Canada Aviation Training Task Proficiency: A Probabilistic Approach. William C. Rankin, William C. McDaniel, US Navy Training Analysis and Evaluation Group, Orlando, FL Air Force Civilian Promotion Appraisal System Development. R. Bruce Gould, Air Force Human Resources Laboratory, Brooks AFB, TX Performance Criteria Development for Army Field Recruiters. Timothy W. Elig, Newell Kent Eaton, Paul A. Gade, US Army Research Institute, Alexandria, VA	Dimensions of Unit Performance I Paper Session; Chair: R. L. Palmer, ARI Validity and Reliability of the Organizational Assessment Package. MAJ Lawrence O. Short, LTC David A. Wilkerson, Leadership and Management Development Center, Maxwell AFB, AL The Measurement of Morale. Melvin J. Kimmel, Francis E. O'Mara, US Army Research Institute, Alexandria, VA Battalion Effectiveness and Command Climate. Francis E. O'Mara, US Army Research Institute, Alexandria, VA The Commander's Unit Analysis Profile (CUAP). R. L. Palmer, US Army Research Institute, Fort Hood, TX
1000	Coffee Break	Coffee Break	Coffee Break	Coffee Break
1030	Adaptive Testing II Paper Session, Chair: M. A. Fischl, ARI Item Response Time as a Basis for Ability and Difficulty Measures. Werner Birke, German Armed Forces Psychological Services Research Institute, Bonn, Federal Republic of Germany Prediction of Boot Camp Attrition: IRT vs. Number-Right Scoring. Robert L. Frey, Jr., US Coast Guard Headquarters, Washington, DC Computerized Optimal Weight Scoring (COWS): A Comparison of Three Procedures. Steven Gorman, Annandale, VA Computerized Adaptive Testing: From Psychometrics to Systems. Paul R. Croll, Office of Personnel Management, Washington, DC	CODAP II Paper Session; Chair: Paul Gade, ARI A Pilot Test of the Feasibility of Implementing CODAP for the Nuclear Power Industry. Jerry B. Hale, Institute of Nuclear Power Operations, Atlanta, GA Interrater Reliability: The Development of an Automated Analysis Tool. Michael R. Staley, Johnny J. Weissmuller, Air Force Human Resources Laboratory, Brooks AFB, TX CODAP: New Techniques to Improve Job-Type Identification and Definition. William J. Phalen, Air Force Human Resources Laboratory, Brooks AFB, TX	Curriculum Development Panel: Course Development and Evaluation Procedures for the Combined Arms and Services Staff School. Chair: LTC Joe Stillman, Command and General Staff College, Fort Leavenworth, KS Michael Anderson, Dennis De Fraia, Ralph Ekwall, Command and General Staff College, Fort Leavenworth, KS	Dimensions of Unit Performance II Paper Session; Chair: Susan E. Kerner, ARI Relationship between Attitudes and Leadership Style: A Policy Capturing Approach. MAJ Jerome Adams, CPT John Richards, CPT Terry Fullerton, US Military Academy, West Point, NY The "New" Officer Evaluation Reporting System. Richard D. Doorley, US Army Military Personnel Center, Alexandria, VA Commanders' Assessment of Unit Effectiveness Measures. Susan E. Kerner, Francis E. O'Mara, US Army Research Institute, Alexandria, VA The Leadership Evaluation and Analysis Program (LEAP): Validity and Future Directions. Thomas D. Affourtit, Interaction Research Institute, Inc., Fairfax, VA
1200	Lunch	Lunch	Lunch	Lunch

1330	Non-Cognitive Measures Paper Session; Chair: Pamela V. Meys, ARI How to Improve the Validity of Personality Questionnaires. Arnold Bohrer, Belgium Armed Forces Psychological Research Section, Brussels, Belgium Biographical Correlates of a Cognitive Abilities Test for Firefighter Selection. Darlene K. Olson, Office of Personnel Management, Washington, DC Self-Assessment in Employee Selection and Placement. Paul van Rijn, Office of Personnel Management, Washington, DC	Occupational Analysis Paper Session; Chair: Douglas Bobko, ARI An Occupational Analysis of Air Force Behavioral Scientists: Preliminary Results. LTC Jimmy L. Mitchell, James Keeth, CPT Linda Wiekhorst, USAF Occupational Measurement Center, Randolph AFB, TX Development of a Work Analysis Questionnaire. MAJ Rodger D. Ballentine, J. W. Cunningham, North Carolina State University, Raleigh, NC An Attempt to Validate a Short Job Evaluation Questionnaire. Walter G. Mann, Jr., Office of Personnel Management, Washington, DC	Performance Appraisal III Paper Session; Chair: CPT Edward W. Williams, ARI Computer-Based Skill Qualification Testing: Development and Validation. Christopher G. Koch, Judith A. Englert, Richard E. Vestewig, Honeywell Systems and Research Center, Minneapolis, MN; John T. Larson, US Army Research Institute, Alexandria, VA Computer Assisted Task Selection in US Army SQT Development. Robert H. Smith, US Army Infantry School, Ft. Benning, GA Job Performance Criteria Development in First Term Troops. Lawrence Maneer, CPT Edward W. Williams, Beverly Popelka, US Army Research Institute, Alexandria, VA	Issues in Military Manning I Paper Session; Chair: Arthur Marcus, ARI Revitalizing the Selective Service Local Board Structure through ADP Technology. Fritz Oelrich, Bruce Bennett, Selective Service System, Washington, DC "Traffic Cop"-Automated Prediction and Control of Registrants to AFEES. B. Michael Berger, Robert Roll, Selective Service System, Washington, DC Effects of Job Satisfaction on Air Force Enlistee Retention. Kent Finstuen, Charles N. Weaver, Air Force Human Resources Laboratory, Brooks AFB, TX
1500	Coffee Break	Coffee Break	Coffee Break	Coffee Break
1515	Predicting Success in Foreign Language Training Paper Session; Chair: Richard P. Kern, ARI Shortening of Defense Language Aptitude Battery. John W. Thain, Defense Language Institute, Presidio of Monterey, CA	Occupational Analysis (con't) Paper Session (con't); Chair: Douglas Bobko, ARI Work Characteristics - A Task-Based, Benchmark Approach. Walter E. Driskill, James B. Keeth, MAJ Frank C. Gentner, USAF Occupational Management Center, Randolph AFB, TX CODAP Users' Meeting; Chair: CDR J. K. Todaro, Executive Agent for Joint Task Analysis Support, and Navy Occupational Development and Analysis Center (NODAC), Washington, DC An informal meeting for all interested in maintenance or enhancement of the CODAP System.	Skill Qualification Testing Panel: Skill Qualification Testing - An Evolving System. Chair: Stanley F. Bolin, US Army Research Institute, Alexandria, VA Robert Eastman, James Sova, US Army Training Support Center, Ft. Eustis, VA; Joan Harman, John Kessler, Douglas Macpherson, US Army Research Institute, Alexandria, VA	Dimensions of Unit Performance III Panel: Analysis of the Impact of Organizational Effectiveness in the Army. Chair: L. W. Oliver, US Army Research Institute, Alexandria, VA U.S. James, Arthur Young Co., Washington, DC; James E. Mietus, US Army Research Institute, Alexandria, Va.
1545	Early Identification of Academic Difficulties in Foreign Language Training. MAJ Allen N. Knox, Defense Language Institute, Presidio of Monterey, CA			
1615	Predicting Officer Success Paper Session; Chair: M. A. Fischl, ARI Academic Preparation, Aptitude Measures and Army Officer Performance. Arthur C. F. Gilbert, Nancy K. Tulloh, US Army Research Institute, Alexandria, VA Army Officer Success in the Different Types of Career Branches. Arthur C. F. Gilbert, Nancy K. Tulloh, US Army Research Institute, Alexandria, VA			

THURSDAY, 29 OCTOBER 1981

	SELECTION AND PSYCHOMETRICS (Capital View Ballroom East)	SPECIAL TOPICS (Capital View Ballroom West)	TRAINING AND PROFICIENCY MEASUREMENT (Terrace I)	ORGANIZATIONAL FACTORS (Terrace II)
0800	Test Development Paper Session; Chair: Richard A. Kass, ARI Development of Enlistment Screening Test (EST) Forms A and B. John J. Mathews, Malcolm J. Ree, Air Force Human Resources Laboratory, Brooks AFB, TX Enlisted Screening Test to Predict Army Area Aptitude Composite Scores. Richard A. Kass, Mary M. Weltin, US Army Research Institute, Alexandria, VA Applying Linguistic Theory to Job-Related Verbal Tests. Alicia Kenaston, Office of Personnel Management, Washington, DC T&E Rating with Objectively Scored Applicant Appraisal Questionnaire. Thomas J. Lyons, Office of Personnel Management, Washington, DC	Systems Development and Evaluation I Panel: Psychological Screening for Weapons Use Suitability. Chair: Samuel E. Krug, Institute for Personality and Ability Testing, Champaign, IL Gary Behrens, Institute for Personality and Ability Testing, Champaign, IL; Robert Palese, Frank Wina, CDR John Williams, USCG, Governor's Island, NY	Program Evaluation I Paper Session; Chair: Bob G. Witmer, ARI Effects of Cognitive Style and Cognitive Dissonance on Students' Course Evaluations. Richard Pigeon, Department of National Defense, Ottawa, Ontario, Canada. Cavalry Fighting Vehicle (CFV) Training Effectiveness Analysis. Walter Butler, Ronald B. Cooper, US Army TRADOC Systems Analysis Activity, White Sands Missile Range, NM A Systematic Approach to Training Program Evaluation. Bob G. Witmer, D. M. Kristiansen, US Army Research Institute, Ft. Knox, KY Skill and Grade Progression of Retrainees Versus Non-Retrainees. Mary J. Skinner, Air Force Human Resources Laboratory, Brooks AFB, TX	Literacy and Readability I Panel: Readability of Materials Chair: 1LT R. Eric Duncan, USAF Occupational Measurement Center, Randolph AFB, TX Richard Braby, J. Peter Kincaid, US Navy Training Analysis and Evaluation Group, Orlando, FL; Lydia Hooke, Tom Sticht, HumRRO, Alexandria, Va.; Hendrick W. Ruck, Nancy A. Thompson, Air Force Human Resources Laboratory; Ron Tar, US Army Training Development Center, Ft. Monroe, VA
1000	Coffee Break	Coffee Break	Coffee Break	Coffee Break
1010	Pilot Selection Paper Session; Chair: John A. Dohme, ARI Royal Air Force Flying Selection. I.G. Fiddurdale, C. C. Elshaw, Air Force Human Resources Laboratory Williams AFB, AZ A Simulated Aircraft Landing Test as a Pilot Selection Test. CPT Pierre Leeard, Canadian Forces Personnel Applied Research Unit, Willowdale, Ontario, Canada The Personality Research Form (PRF) as a Prediction for Success in Pilot Training. [illegible] erdyke, Canadian Forces Personnel Applied Research Unit, Willowdale, Ontario, Canada The Fairness of the Army Flight Aptitude Selection Test. John A. Dohme, US Army Research Institute, Ft. Rucker, AL	Systems Development and Evaluation II Paper Session; Chair: Gary G. Sarli, ARI Human Factors Evaluation of Two Division Air Defense Gun Systems. Gary G. Sarli, Richard J. Carter, US Army Research Institute, Ft. Bliss, TX Soldier/Machine Interface Evaluation of Army Developmental Systems. James C. Geddie, US Army Human Engineering Laboratory, Aberdeen, MD Job Related Measurement of Reading Ability. Sandra S. Payne, Office of Personnel Management, Washington, DC	Program Evaluation II Paper Session; Chair: Stanley F. Bolin, ARI The Strategic Weapon System Training Program. I: Description. F. B. Braun, Arlington, VA The Strategic Weapon System Training Program. II: Executive Steering Group's Role--S.W.S. Personnel and Training Evaluation Program. Edward Hobson, Arlington, VA Evaluating Training Using a Confidence Interval Technique. John L. Derhammer, US Army Infantry School Ft. Benning, GA	Literacy and Readability II Panel: Literacy: Assessment and Remediation Chair: 1LT R. Eric Duncan, USAF Occupational Measurement Center, Randolph AFB, TX John Claudy, American Institutes for Research, Palo Alto, CA; M. A. Fischl, Richard P. Kern, US Army Research Institute, Alexandria, VA; Robert Wisher, Navy Personnel R&D Center, San Diego, CA
1230	Lunch	Lunch	Lunch	Lunch

1330

Enlisted Success

Paper Session; Chair: Robert Pleban, ARI

Selection Predictors in the British Infantry. Carol Heaton, Army Personnel Research Establishment, Farnborough, Hants, United Kingdom

Use of the ASVAB as a Predictor of Early Discharge. Edward Kemery, Robert Pleban, Beverly Williams, Frederick Dyer, US Army Research Institute, Ft. Benning, GA

Predictors of Success in Basic Enlisted Submarine School. LCDR Earl H. Potter, III, US Coast Guard Academy, New London, CT

Computer Applications I

Paper Session; Chair: Susan E. Kerner, ARI

An Innovative Approach to Data Capture in Automated Assessment. MAJ John F. Taylor, Army School of Training Support, Beaconsfield, Buckinghampshire, England

Simulation of Command Group Operations: An Evaluative Report. MAJ John F. Taylor, Army School of Training Support, Beaconsfield, Buckinghampshire, England

AREIS: A Computer-Based Educational Counseling System for the Army. Melissa Berkowitz, US Army Research Institute, Alexandria, VA.

Achievement Measurement

Panel: Theory-Based Approaches to Constructing and Calibrating Achievement Measures.

Chair: Harry O'Neil, US Army Research Institute, Alexandria, VA

Eva L. Baker, Bruce Choppin, Edys Quellmalz, Center for the Study of Evaluation, UCLA Graduate School of Education, Los Angeles, CA

Officer Performance I

Paper Session; Chair: Michael G. Rumsey, ARI

Race Influences on Peer Ratings in ROTC Training Platoons. Michael G. Rumsey, US Army Research Institute, Alexandria, VA

Twenty Years Later: A Follow-Up of the Officer Evaluation Center. Fae V. Hays, Frederick N. Dyer, US Army Research Institute, Ft. Benning, GA

An Intense Training Program that Doubles as a Personnel Assessment Center. CDR John B. Washburn, Chief of Naval Education and Training, Pensacola, FL

The Honor Code: The Optimum Criterion of Honor? COL Clark L. Hosmer, USAF (Ret.), Shalimar, FL

1500 Coffee Break | Coffee Break | Coffee Break

1515

Test Validity and Usefulness

Paper Session, Chair: John R. Mietus, ARI

Validation of RCMP Selection Procedures. Leonard Wevrick, Ministry of the Solicitor General, Ottawa, Ontario, Canada

Empirical Bayes Assessment of Differential Validity. Henry T. Braun, Douglas R. Jones, Educational Testing Service, Princeton, NJ

The Problem of Range Restriction in Test Validation. Robert C. McKenzie, US Office of Personnel Management, Washington, DC

The Usefulness of Selection Tests. Albert E. Beaton, John L. Barone, Educational Testing Service, Princeton, NJ

Computer Applications II

Paper Session; Chair: William B. Camm, ARI

Improving Acceptance of Automated Counseling Procedures. James M. Johnson, Illinois Institute of Technology, Chicago, IL

Privacy Act and the Data Base: Implementation of the Privacy Act. William B. Camm, US Army Research Institute, Alexandria, VA

Videodisc Technology

Paper Session; Chair: Frederick N. Dyer, ARI

Exploration of Instructional Applications of Videodisc in the Marine Corps. William H. Greenup, Marine Corps Development and Education Command, Quantico, VA; Steven Levin, Interactive Television Company, Arlington, VA.

Interactive Video for Interpersonal Skills Training. Frederick N. Dyer, James E. Schroeder, US Army Research Institute, Ft. Benning, GA; Paul Curray, Daniel P. Gillotti, Edward W. Youngling, Litton Mellonics, Ft. Benning, GA.

1830	Reception Cash Bar	
2000	BANQUET	
	INVITED ADDRESS:	"New Horizons in Manpower," MG Jack I. Posner, US Air Force (Ret.), General Research Corporation, McLean, VA
	Music: DC National Guard 257th US Army Band	

FRIDAY, 30 OCTOBER 1981

	SELECTION AND PSYCHOMETRICS (Capital View Ballroom)	ORGANIZATIONAL FACTORS (Terrace II)
0900	Special Topics Paper Session; Chair: Francis E. O'Mara, ARI <u>Current Directions in Selection Research for the Royal Navy</u>. B. T. Dodd, Royal Navy School of Educational and Training Technology, Portsmouth, England <u>Reading Ability, Readability, Motivation and Test Validity</u>. Clay V. Brittain, Mary M. Brittain, Williamsburg, VA <u>NROTC Joins the BIG Test Movement</u>. Andrew N. Dow, Office of the Chief of Naval Education and Training, Pensacola, FL	Issues in Military Manning II Panel: <u>Military Compensation: A New Look at an Old Challenge</u>. Chair: Linda Pappas Hale, Hay Associates Washington, DC CPT Thomas Hale, USN Office of Chief of Naval Operations, Washington, DC; Peter Oglobin, Office of the Secretary of Defense/MPP, Washington, DC
1030	Coffee Break (Capital View Ballroom)	

1045 PLENARY SESSION AND CONFERENCE CLOSING
(Capital View Ballroom)

Current Status and Milestones for MTA 25th Anniversary Publication

Raymond O. Waldkoetter, US Army Research Institute, Fort Sill, OK

Closing Remarks

COL L. Neale Cosby, Commander, US Army Research Institute, Alexandria, VA

Contributed Papers

Copies of individual papers presented are arranged alphabetically in the PAPER PRESENTATIONS section of the Proceedings, by principal author. Papers which were presented as part of a panel session are to be found in the PANEL DISCUSSIONS section of the Proceedings, which are alphabetical by chair. The Conference Program at the front of the Proceedings, in conjunction with the Author Index, which follows, may be useful in locating a specific paper of interest.

No attempt has been made to edit papers. The views expressed in them are those of their authors, and not necessarily those of the organizations which they represent. Correctness is the sole responsibility of the authors.

AUTHOR INDEX

Panel Discussions

KEYNOTE ADDRESS

LTG Paul F. Gorman
Assistant to the Chairman, Joint Chiefs of Staff

A Rumination On Wars Ahead

Remarks to the Military Testing Association
Washington, D.C.

26 October 1981

Lt. Gen. P. F. Gorman, U.S. Army

Slide 1

I will be talking forecasting with you, but let me make it clear that mine is not a Schweitzerian prediction of inevitable drift towards war, but rather the sort of concern that all of us in the Pentagon have to direct toward the future in order to do our job. Bob Schweitzer aside, I would have you note that Nick "the Greek" Snyder quotes odds of ten to one that there will be a war involving our forces within the coming decade. As the present Secretary of Defense has stated, it is the business of his Department to prepare for future wars. Wars and rumors of war are problems with which we must seriously contend. Let me begin by reading you a quotation from the Defense Guidance, which as some of you may know, is a document wherein the Secretary of Defense instructs the members of his own staff and the Joint Chiefs of Staff on how he wants them to prepare for the future. The time frame for planning in the current draft Defense Guidance is the fiscal years 1984 through 1988. Thereby you can see we are already conceptually far into the future. I will quote from a section of the guidance which deals with our using to advantage those elements of the American system wherein we might be able to enjoy some edge over the competition:

> "An inherent advantage of our system is the ability to combine managerial skills and technology to solve difficult problems. Getting serious about competing with the Soviets will require that the U.S. use its advantages to develop an overall strategy with several areas of primary emphasis: (1) developing first-rate weapons and eliminating some of the reliabilty problems we have recently experienced;

(2) expanding the competition into other areas where the Soviets find it difficult to compete; (3) designing innovative forces and operational tactics and procedures that act against the Soviet control system and frustrate, disrupt, and defeat Soviet attempts to achieve political/military objectives; and (4) developing cost imposing strategies whose basic aim is to obsolesce past Soviet investments."

Now, that guidance is a tall order. The purpose of my discussion is to present considerations which someone in your line of work, as well as mine, might bring to bear upon an attempt to comply, either from '84 to '88, or in the years beyond. Let me start by laying out for you some of the trends in warfare of which I am aware, and which you may want to take into account.

Slide 2

First of all, the long sweeping line ascending from the year 1860, the time of our Civil War, to the 1990's in the upper right-hand side of the chart traces the amount of area that a battalion--a group of 600-800 men--was expected to control given the firepower and mobility means available to them at any given point in time. The fundamental reason for having military forces at all is to control land and people, and the trends are all in the direction of doing more militarily with fewer soldiers. As you can see, I have extrapolated, like most futurists, control trends beyond the present. I would also call to your attention that, like most futurists, I have used a logarithmic scale on the ordinate. The bars graph firepower, for which the measure of merit is the pounds of projectiles that a division, a large assemblage of battalions, could hurl at an enemy in the course of an hour, all guns firing maximum rate, for each soldier engaged. You will note that the bar graph heights do not keep pace with the extention of the area of control, and that lack of direct correlation is a function of both the superior mobility means that we have made available to our land warfare organizations--large quantities of tracked vehicles, and helicopters, plus advanced communications--as well as the efficiency of modern ordnance--one need not project as much ordnance in terms of poundage for greater lethality, given more efficient conventional munitions, and the so-called "smart" weaponry in the force. Finally, on the slide, slashing down diagonally over the trends just described, is a measure that I have called "division dispersion." Here I have taken an amount of the forward edge of the battle area (FEBA) that the division is responsible for in normal disposition, and computed a measure of men per kilometer of FEBA. And as you can see, those trends plunge linearly downward, and my projections suggest to you that they will keep going down through the 1990's. What that line tells you is that we have multiplied the ability of each soldier on the battlefield, or that the density of men on the battlefield will be going down, and going down dramatically, continuing the trends that we have observed as technology has been applied to battle in recent years. Each soldier in battle will count for more toward accomplishing the basic mission of control of the earth's surface.

Slide 3

I believe that it is possible, looking into the future, to conclude that, almost certainly, a technolgically advanced combatant in future warfare will be able to see all elements of an opposing forces in real time, and will have at his disposal firepower means for reaching out to strike throughout the depth of the opponent's war-waging apparatus from his theater forces all the way back to his strategic reserves. Some naval officers have found it reasonable to say, vis-a-vis naval warfare, that it will be difficult if not impossible to steam around the seas with forces centered on a large-decked carrier, with protective rings of specialized air and submarine defense ships around that carrier. Some air officers have found it possible to say that we will have to find alternatives to the operation of air forces from large fixed airfields, whereon aircraft are processed for high sortie generation rates on something like an assembly line basis. I can assert for land warfare that the day will soon be gone when massed formations of armored vehicles will be able to swarm over the surface of the earth, trailing behind elaborate logistic tails. Instead we are going to have to move toward something like "distributed force," meaning that in order to provide protection we will have to disperse more broadly and thoroughly than ever before, and thus confuse the enemy as to which elements of the target array before them are particularly significant as threat. Our tactical dispositions will have to confront our foe with a large complex of target elements, each of which is potentially able to deliver punishing firepower, and each element of which could be capable of developing the intelligence requisite to the accurate delivery of that firepower.

Now, there are enormous impediments both technological and cultural to achieving such a capability. But I am convinced that that nation who is first able to achieve the desiderata that I have sketched will exert an enormous superiority over potential adversaries, and I suggest that the excerpt of the Defense Guidance that I just read you is quite right: it would be important for the United States, and any other nations of the free world who wish to assist in the competition with the Soviet Union, to bend their efforts to field first-rate weapons, and invent new tactics and techniques for using them.

Slide 4

Here is a description of future land warfare which I have drawn from a presentation made to the Secretary of Defense just this past summer by the Defense Science Board, which as many of you may know is a group of prestigious scientists who provide advice to the Secretary on problems of particular significance. This summer he asked several DSB Panels to look into the future of warfare, and to advise him what technologies might be relevant to the Department of Defense as it sought to implement the guidance that I read at the outset of my remarks. Here you can see the notion of "distributed force" in an athletic analog, a way of describing warfare which is always more comfortable for Americans than most.

Slide 5

But such amorphous warfare will generate a series of new needs or requirements, of which the Defense Science Board cited these. In other words, these needs are problems that technology has to solve in order for us to have the capability to compete in land warfare as just described. The red signals on the right indicate needs which I submit are as much sociological as they are technological, as much a demand for cultural evolution as they are matters for advanced science and materials. And it is to the needs for cultural leaps that I will eventually direct our attention here today.

But before I do so, let me quickly walk you through some of the technological responses to these needs that the DSB Panel cited for the Secretrary of Defense.

Slide 6

In the first place, they pointed out the United States faces a very difficult choice, as indicated here: Shall we, as many advocate, settle for less complicated, less technoligically advanced weapon systems, systems of only sufficient capability, and, since their price will be lower, thereby assure that we can purchase more weapons? Or shall we rather take the course advocated by the Defense Guidance and reach for excellence, knowing full well that, if we do, we thereby expose ourselves to high expense, and concomitantly have to settle for relatively few of those weapon systems? All pertinent trends are depressing:

Slide 7

Here is a slide used by Norman Augustine, Chairman of the Defense Science Board. He calls it "Calvin Coolidge's Revenge" because President Coolidge, when presented military budget for 1928, is reported to have said, anent an item of some $25,000 for the purchase of a squadron of airplanes, "Why can't we buy just one airplane and let the aviators take turns?" Mr. Augustine' chart makes it evident that by the year 2054 the entire Defense Budget of the United States will be able to purchase just one airplane, so we will have to let the Navy use it 3-1/2 days each week and the Air Force the remainder. And, of course, by the year 2100, it predicts that the entire gross national product of the United States will buy just one airplane, and we will have unification of the services at last.

Slide 8

Not only are cost trends up, but as the systems become more expensive, the trends are to buy fewer of them: whether it's carriers of the NIMITZ class, down on the bottom right, the AWACS, or various advanced air defense systems, the more expensive the system is, the fewer that you will see in inventory. Hence, we are indeed very much in a numbers quandry.

Slide 9

Moreover, as the complexity and costs of systems increase, reliability seems to decrease. Here is a plot Mr. Augustine has put together comparing costs of various items of recent aviation electronic equipment to hours between failure. This chart, he says, illustrates that if you are willing to pay enough for a given avionic apparatus, you can guarantee that it won't work at all.

Slide 10

However, this summer the Defense Science Board, after examining just such trends and propositions, came to this rather opposite conclusion. Why?

Slide 11

Well, among other things they looked at pairings of advanced systems with older counterparts. For example, here the F-14 and the F-4J, a Naval fighter new and old; Patriot and Improved Hawk, SAM air defense systems new and old; the F-15 and F-4E, Air Force fighter aircraft, new and old; and the M-1 and M-60 A3, Army tanks, new and old. The values on the ordinate express effectiveness in terms of ratio of kill potential, new to old (you'll just have to take the DSB Panel's word that these are reduced some kind of common base for comparison). And on the abscissa is plotted the ratio of acquisition costs, new to old. The parity line represents the break even point, system for system. As you can see, the very much increased effectiveness of the new systems drives all of these pairings well up into a win area of much better than parity: 3, 4, 6 to 1.

Slide 12

The same thing is true in terms of support costs, which get at the issue of keeping systems available or ready for use in the force. Support costs include the costs of parts, training the operator and the maintainers to keep the system operating, and the costs of requisite basing. Again, the picture here is favorable for the new systems. So it is possible, according to the DSB, to build into new systems both high availability rates and very high effectiveness, and thereby to compensate for reduced numbers of systems. In fact, as you can see, with effectiveness ratios of 5 to 1 or so over systems they are replacing, one can operate with significantly fewer elements in the force.

Slide 13

Indeed, one DSB panel made for the Secretary of Defense this kind of a case for high performance or high technology systems. We might buy twice as many "half-performance systems," but the panel noted that the Soviets are building a very large, very high performance force, against which a smaller lower performance force would do very badly. Half-performance systems could not defeat Soviet systems built on advanced technology, such as we see in their tanks or in their titanium hull submarine. Moreover, high performance systems give us for the first time the ability to operate at night, in adverse weather, and under the electronic warfare conditions which will be typical of the future. Moreover, we can get more "sorties" out of such equipment than has been possible in the past. (Incidentally, for those of you who have been following the dialogue in Congress over whether some particular fighter aircraft is better than the one it is replacing, should note that very frequently critics of our latest airplanes will contend that the sortie rate of the new is below that of a plane we had in World War II or Korea. You have to remember that sortie rates for the present peacetime force are programmed by flying hours and parts, and we don't fly these airplanes more often because we didn't have funds to do so. But our new fighters, in tests under field maneuver conditions in Europe and elsewhere, have demonstrated a capability to produce sortie rates two and three times what we have had in wars past.) Perhaps more importantly, to accept lower performance systems would be simply to accept higher, avoidable U.S. casualties. And if we went the low-performance route, we would need to increase our intake of military manpower overall not only to compensate for losses, but to operate the increased numbers of systems in the force. True, these could conceivably be less demanding systems in terms of skill requirements.

But higher numbers of systems would increase our already high support costs, and--the point needs to be made again and again--human costs, manpower costs, will dominate the life cycle cost of virtually every system that the Department of Defense presently has under procurement or under research and development. And then finally, of course, maintenance support for larger numbers of systems in the force would increase the strain on a force which is already stressed in providing for its large logistic tail, its training base, its management, housing for dependents, personnel services, and headquarters overhead. The last line on the slide, basing, makes the additional point that whether you are talking about aircraft carrier decks, or airfields, or tank parks, we today have a constrained basing system for the American armed forces; putting a lot more systems into the field could create problems of basing for us.

Slide 14

Well, what high technology should we reach for to acquire high performance systems, pursuant to the Defense Guidance? Here is a list of technologies which the Defense Science Board believes could make an order of magnitude difference in our capability to meet the exegencies of future warfare. What you are looking at are results of a Delphi technique, in which the panelists assigned measures the merit to various technologies in order to assess relative opportunity and relative risks, from which they derived the numerical rating on the right, a function of both opportunity and risk.

I am going to talk about a number of these particular technologies in a moment, but note that overall the Defense Science Board commended some 17 technologies to the Secretary. I have deliberately avoided having to discuss some of the more arcane--such as "high density monolithic focal plan arrays"--in order to focus on some of the technologies which I find, interestingly enough, are both at the top of the list and prominent for their sociological or cultural dimensions.

Slide 15

For example, this one. On the right you see the investment posture in the Defense Research and Development budget at the moment, and elsewhere on the slide, an indication of what Defense might do by way of applications: what kinds of opportunities the technology could provide. I would have you note that the Defense Science Board panel identified computer software as the highest risk undertaking among all Defense development undertakings.

Slide 16

This chart, rather startling to some, indicates the problem that we face unless we are able to find ways of producing computer software more efficiently than we have in the past. Without some technological intervention, we simply are not going to be able to exploit the promise of smaller, smarter processors. The United States is facing very severe shortages of system analysts, computer scienti and programmers. The research into advanced software methods to which I just alluded, does not hold out promise for relief for many years, but it is hopeful, and we should pursue it.

Slide 17

And then you have a possibility of improving the inherent capability of materiel with respect to its availability or reliability in the field. In commending this particular technology to the Secretary of Defense, the Defense Science Board stated as follows:

> "Reliability standards must be raised significantly--the technology to support such increases is available--the adherence to these standards the first time around is the most economical approach in the long run. Front-end costs will be higher....Times and funds must be programmed in the development cycle to accommodate necessary redesign iterations after test and before Initial Operating Capability for critical reliability, maintainability and producibility problems (as well as performance problems)."

Slide 18

And then, lastly, for the moment at least, this particular technology. A number of recent publications have made it evident that by the end of the decade of the 1980's there will be available machines, computers or processors, which will be comparable in size and in weight to the human brain, and will have comparable requirements for input and energy, and which will have quite comparable capabilities for output. Now what that means for the weapon system developer is obvious.

Slide 19

Here is a technology which was not cited by the DSB, but one I picked to highlight further the human dimension of the problem. This is a mechanism for propelling a projectile using electro-motive vice chemical energy, exploiting the so called Lorentz force. By the end of this calendar year, 1981, in the Westinghouse R&D Center at Pittsburgh, Pennsylvania, the Defense Advanced Research Projects Agency and the Army's Development and Readiness Command hopes to have operational a laboratory device which will be able to propel a 300-gram mass to a speed of 3 kilometers per second, yielding a muzzle energy on the order of 1.35 megajoules. In effect, this laboratory device will shoot a bullet about 3 times faster than present rifles or tank guns, and opens a whole new realm of physical possibilities with respect to antiaircraft guns, tank guns, and that sort of thing--Tom Swift's Electric Gun at last. Such weapons, if they ever become practical, will also propel us into a whole new realm of difficulties with that most intransigent, stultified, subculture within military sociology, the artillery.

Slide 20

In 1977 when I took command of the 8th Division in Germany, I found this state of affairs in my division artillery. I found that my artillerymen adhered to their manpower-intensive ways of doing business as assiduously as their forefathers had clung to horsepower. Application of high school physics--levers and pulleys, inclined planes--could have bettered the situation, let alone buying materiel handling equipment, or looking to robotics to solve the problem. Let me read to you from a recent Army War College pamphlet on the Army in the year 2000.

> "In considering technology the Army must look introspectively at its ability to use advanced technology and its past performance in this area. For over 20 years there has been the technological capability to have a Howitzer that could be electronically laid (directed), fuzes automatically set, rounds automatically rammed, muzzle velocity (for future corrections) electronically measured and firing data electronically computed from an electronic sensing. The actual condition is that there are many artillery commanders taking great pride in the fact that they never fire their Howitzers using only the FADAC (a very old computer which is dependent on mobile generators usually in short supply). These commanders insist on checking the FADAC by manual means or they check the manual using the FADAC. One could imagine the confusion resulting from the introduction of the modern artillery systems which we should have."

I can attest to that. I sent back from Germany a young officer to the Army Materiel Systems Analysis Agency where, working with ballistic experts, he developed a chip for a Texas Instruments programmable hand-held calculator incorporating the firing tables for our medium howitzers. The device proved both reliable and quick. But I can vividly recall walking into a fire direction center of a battalion firing at Grafenwohr to find three tiers of firing data calculators in operation. On tier one of a football-bleacher like set-up there were three plotters producing data for the guns using the graphic techniques that had been used in World War I and World War II. On tier two were not one but two FADAC operators checking the graphic data. And in the back at the top of the pyramid was a lad with the Texas Instrument device, invariably producing his data faster and as accurately as anybody else, but whose data would not be independently accepted by any self-respecting fire control officer, you may rest assured.

What I am suggesting is that a technology like electric guns, which could eliminate chemical energy as a way of propelling projectiles, could also eliminate an enormous amount of that manpower intensive logistic tail to wnich I alluded at the outset, and may make it possible for us to have a genuinely distributed force. In future wars we will have to dispense with specialization. I do not believe that it will be possible to maintain through the 1990's an arm dedicated, as we now dedicate the artillery, exclusively to the delivery of indirect firepower. I think all elements of the force are going to have to be capable of contributing to both direct and indirect firepower, to anti-aircraft and anti tank defenses, and to reconnaissance. Hence, our present specialized, manpower-intensive artillery has got to yield place to multi-purpose weapons systems.

Slide 21

This is a cartoon suggesting such a future weapon, by William Coulter of the Washington Post. It is not, I assure you, genuinely classified. It depicts a device which I admire if only because it operates by ingesting its own technical documentation. I believe, as the data I am about to show will illustrate, that such a feature in a weapon may be the only hope for the Army of the future. But I want to use this slide to make a more serious point about the dilemma that we face: here the artist caricatures one of those omnipotent machines we might have to develop. I would say that the principal obstacle to our fielding to such a machine is the fellow portrayed sitting on the seat there on the left: the commander, operator, maintainer of the device. Here again, a quotation from a Defense Science Board report to the Secretary this summer:

> "The division of tasks between the man and the machine becomes increasingly critical in two dimensions. First, there is the problem of personnel skill potential (quality). Average reading levels and

aptitudes come into play. Secondly, there is the complexity issue. To the extent that complexity can be engineered away from the man-machine interface so much the better, if it can be afforded, and if it is not translated into insoluable problems somewhere back in their maintenance sequence. More human research and man-machine technology development is required. The current problems with Built-In Test Equipment (BITE) illustrate the doubtful state of the art. The division of maintenance tasks between the diagnostic equipment and the mechanic or repairman has been tilted too far toward the machine and they have generally failed to live up to their advanced billing. In the meanwhile, the people people must interact realistically with the engineers at the outset. This is an art not yet fully developed. Testing at the man-machine interface must be conducted and room for corrective design interactions provided in the development program."

Slide 22

I would be remiss if I did not immediately, of course, acknowledge, that in addition to the issue of the quality of the men manning weapon systems, we have a quantity problem in this country. The Soviet Union faces a similar quantity problem, and so too, I believe, do most developed countries.

Slide 23

But we also have a quality problem, and it may be societal in scope, as illustrated by the decline in Scholastic Aptitude Test results nationwide.

Slide 24

But even more disturbingly, recognizing that we are in a serious competition with the USSR, our high school graduates, compared with Soviet high school graduates of the past decade, have had far less disciplining in mathematics, science and other technologically supportive subjects in the course of their schooling.

Slide 25

I quote here from Dr. Isaak Wirzup, Professor of Mathematics at the University of Chicago, and Director of the East European Survey of Mathematical Literature for the National Science Foundation, and also the NSF's Director of nical documentation per tank, and now fielding over 40,000 pages per tank. It is easy to imagine what that means in terms of complexity for the mechanic who has to be able to find his path into his technical documentation for fault isolation, repair, replacement, etc. What such a complication does in terms of slowing the rate of repair, and the reliability of the repair is predictable, and the predictable is happening already in some of our current higher-technology weapons systems.

Slide 29

Here is a flow diagram showing the maintenance process for the anti-aircraft missile, the Improved Hawk. Up at the top you see the Hawk unit's own maintenance, and across the bottom the direct support maintenance performed exogenous to the Hawk unit. The highlighted figures suggest that of parts that were sent from the unit back to the direct support level for repair, 40% when examined were found to be faultless; that is to say, almost half the time the Hawk unit removed parts from the weapons system and sent them away to the direct support maintenance unit on a totally unnecessary trip. Moreover, you can see over on the left, when the parts came out of the direct support maintenance unit and were returned to the Hawk unit, 30% of the time they didn't check out when actually refitted to the system. And up at the top, the box in the center suggests that these mishaps are a function of inaccurate trouble shooting, of inadequacies in the built-in test equipment, and of personnel deficiencies in training, quantity or experience.

Slide 30

These data were compared with the I-Hawk experience of the Bundeswehr, at least with respect to the no-evidence-of-failure rate. There was a dramatic difference of over 20 times less maldiagnosis.

its program on Soviet Applications of Computers to Management. In a letter to the National Science Foundation, Dr. Wirzup said this:

> "The Soviet Union's tremendous investment in human resources, unprecedented achievements in the education of the general population, and immense manpower pool in science and technology will have an immeasurable impact on that country's scientific, industrial and military strength. It is my considered opinion that the recent Soviet educational mobilization, although not as spectacular as the launching of the first Sputnik, poses a formidable challenge to the national security of the United States, one that is far more threatening than any in the past and one that will be much more difficult to meet."

Slide 26

I think you are familiar with these data. Regardless of what you may think of their precision, the trends are plain. The population recruited into the Army recently was different qualitatively from the population that was inducted or recruited in the wanning years of conscription in the United States. This is not to say that the Army cannot work successfully with a large median population. However, it is evident that the Army encounters significant difficulties when attempting to assign masses of such individuals to high technology weapon systems. I am not talking about future difficulties; I am talking about problems that are here and with us today. Reference was made earlier to reading ability.

Slide 27

This is what is happening to technical documentation vis-a-vis airplanes. As you can see, the Navy's F-14 fighter is going into the fleet accompanied by some 300,000 pages of technical documentation.

Slide 28

And it is not just airplanes, but tanks: here you can see the Army's starting early in World War II with something less than 1,000 pages of tech-

Slide 31

As you might expect, if an armed force is willing for a technologically advanced system such as the Improved Hawk, to invest individuals with 8-year terms of service, who are all high school graduates, and who have other advantages over what the U.S. Army has been working with on the left, the performance of the weapon system is almost invariably going to look better.

Slide 32

Now, it isn't that we are picking less apt soldiers to man the Improved Hawk as a system. These are Electronic Aptitude scores. As you can see, compared with the total force, the I-Hawk system is getting a pretty good cut of the soldiers of higher electronic aptitudes. And, of course, the I-Hawk is but one of many systems that have a demand for soldiers with these high scores. Part of our problem is that the tests which produced these scores are normative. Part of our problem is that even high scoring soldiers may not be up to the maintenance task to which we have put them. Part of our problem is exactly that the weapon system may be too smart, and we need to engineer it some more in order to drive its maintenance back into routines with which we know the soldiers we have can cope. These are all very, very intractable problems, to handle which the Army of the future, or the Air Force of the future, or the Navy of the future, are going to need the skills of you psychometricians far more then ever in the past.

Slide 33

What to do? I advocate a manpower strategy for the Armed Forces, in complying with the guidance of the Secretary of Defense, which would operate on all three phases or learning regimes of the soldier, sailor or airman. In the past, of course, the services have focused resources on the center phase, labeled here the school, and they will have to keep doing so. They have sort of left it to hope, to chance or to high enlistment bonuses that the Phase 0 product would pan out for them: not a sound approach, given what is happening in SAT scores. We are going to have to do something about finding individuals in Phase 0 who are fit to become stellar performers both in Phase I and Phase II. But it is equally important to devise a plan for Phase II--On the Job. It has been the custom of the Services of the United States to let Phase II, the on-the-job phase, take care of itself, relegating what may be the most important adult educational experience to first-line supervisors largely ill-trained for teaching, and unaware of their responsibility for same. I suggest to you that we can no longer depend upon such hit-or-miss methods if we are going to modernize the force, to bring in large numbers of very technologically advanced weapon systems. I hold that we must launch now a concerted campaign to intervene systematically in Phase II, the on-the-job training phase of development for the soldier, sailor or airman, both to assure functional mastery, and to provide an ability in the force to handle the influx of newer technologies that will pour out of on-going research and development programs.

Slide 34

Is there a technological intervention appropriate here? Yes, you may recall Number 4 from the list of 17 DSB commended technologies, this one. One last time I quote, if I may, from a Defense Science Board report:

> "It should also be the policy of DOD that support will be provided for these high performance systems at a level which will meet peacetime operating and training requirements and which also will provide the base for surging to wartime utilization and sustainment rates. In wartime intense combat periods, and during peacetime 'surge trials' it will be the objective to move actual field availability A_0 close to A_i, intrinsic availability. Specific support program goals should be established at the beginning of development and managed thereafter with the same priority attention and intensity normally accorded to performance. Training support goals should relate to higher standards based on advanced training

technology now becoming available. Advanced job-aids should be designed for simultaneous use in training and on the job. Soon it will be impossible to maintain this kind of technical documentation by conventional generation, distribution, and substitution of paper drawings and text. Digital communication, storage, and display of changes will be required. This whole area should be promoted during the acquisition cycle not only by R&D community, but by personnel specialists and commanders."

Now ladies and gentlemen, it is not clear at this time whether the strategy advocated in the Defense Guidance will enhance the national security of the United States. Both its technological and sociological feasibility are seriously in doubt. The manpower policy challenges posed by that strategy are enormous in their implications. The branchings in the paths ahead are altogether too numerous for easy mapping or classification. It seems crystal clear, however, that most of them involve choices that could more confidently be made were we to have much better information about our manpower than is presently available to commanders and managers in the Department of Defense. Hence, I appeal to you, individually and collectively, to lend us a hand with your skills. Upon your response, the very security of the nation may rest.

ORGANIZATION OF THE
JOINT CHIEFS OF STAFF
IATE OF ADMINISTRATIVE SERVICES

GRAPHICS BRANCH

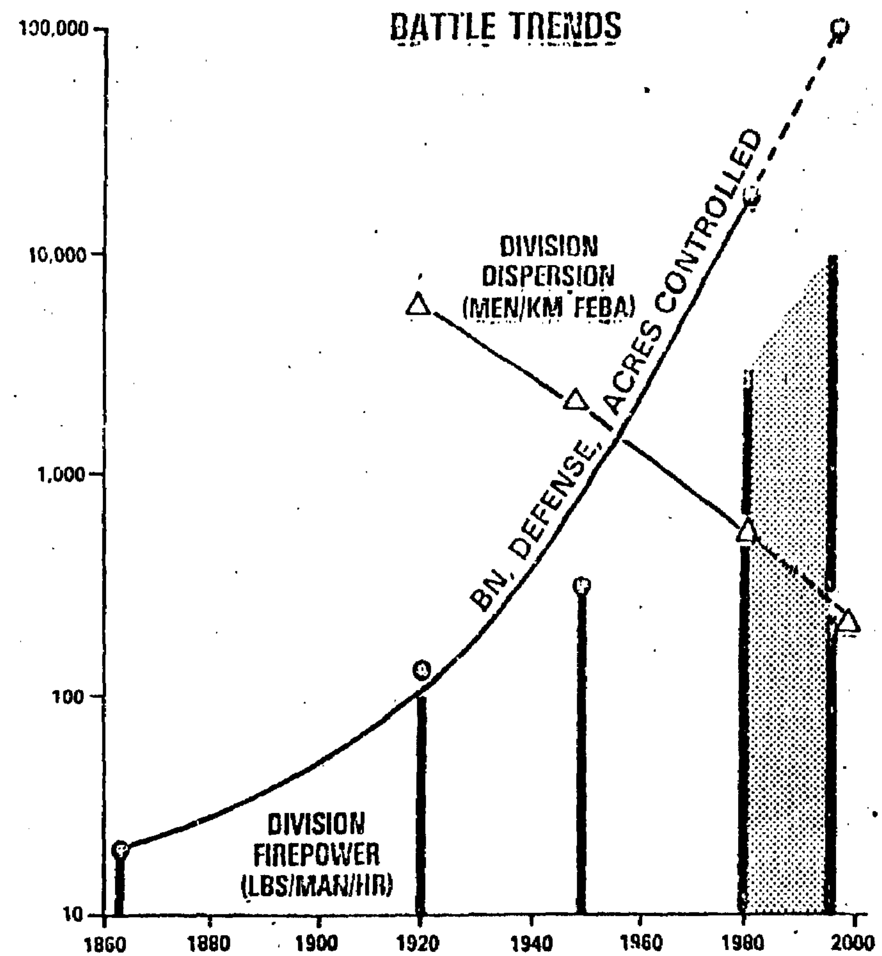
1

A RUMINATION ON WARS AHEAD

5323-168

ORGANIZATION OF THE
JOINT CHIEFS OF STAFF
ATE OF ADMINISTRATIVE SERVICES
GRAPHICS BRANCH

a



5323-160Z

3

NOW	FUTURE
○ LARGE SURFACE COMBATANTS ° RINGS AROUND-CORE DEPLOYMENTS	
○ LARGE FIXED AIRFIELDS ° INDUSTRIAL-MODEL OPERATIONS	○ DISTRIBUTED FORCE
○ LARGE ARMOR/MECHANIZED MASS MANEUVER ° CONVEYOR-BELT LOGISTICS	

5323-168F

4

LAND WARFARE POST-1990

- CONFLICT MORE ANALOGOUS TO SOCCER THAN FOOTBALL
  - CONTINUOUS ACTION VS "DOWNS"
  - OFFENSE AND DEFENSE VS SPECIALISTS
  - FLUID CONDITIONS VS FIXED PLAYS

- MAJOR CONCENTRATION OF TROOPS AND EQUIPMENT MAY NOT BE PRACTICAL
  - TACTICAL NUCLEAR WEAPONS AND WIDE-AREA MUNITIONS
  - SURFACE-TO-SURFACE AND AIR-TO-SURFACE MISSILES
  - PRECISION EMITTER LOCATION

- NEW TRADE-OFFS AMONG MOBILITY, AGILITY, AND FIREPOWER

- THE NEED TO CONDUCT SUSTAINED OPERATIONS (DAYS NOT HOURS)
- THE NEED TO LOCATE AND TRACK MAJOR ENEMY FORCES CONTINUOUSLY
- THE NEED FOR REAL-TIME INFORMATION MANAGEMENT
- THE NEED TO COUNTER ENEMY TARGET ACQUISITION SYSTEMS
- THE NEED TO MORE CLOSELY INTEGRATE THE "EYEBALL AND THE TRIGGER"
- THE NEED FOR SECURE, JAM-RESISTANT, FAIL-SOFT, MOBILE COMMUNICATIONS
- THE NEED FOR DISPERSED, SMALL UNIT CAPABILITY
- THE NEED FOR "TRANSPARENT" COMPLEXITY
- THE NEED FOR EQUIPMENT AVAILABILITY AS WELL AS RELIABILITY
- THE NEED FOR SYSTEMS THAT CAN OPERATE IN EXTREME ENVIRONMENTS (CB/RADIATION/TEMP.)

ORGANIZATION OF THE
JOINT CHIEFS OF STAFF
OF ADMINISTRATIVE SERVICES
GRAPHICS BRANCH

6

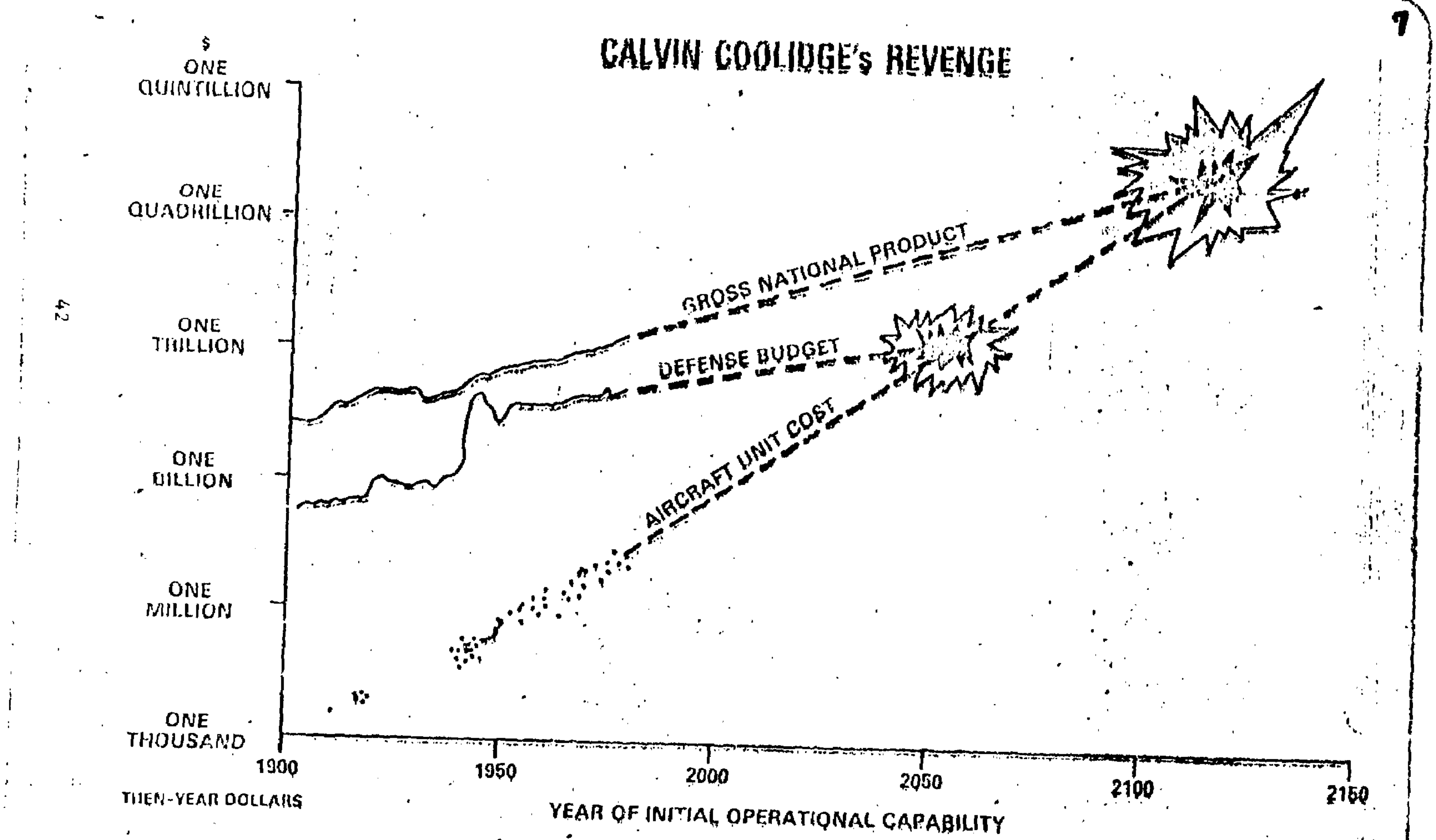
TECHNOLOGICAL RESPONSE?

EXCELLENT	SUFFICIENT
EXPENSIVE	CHEAP
FEW	MANY

5323-168E

FILE NUMBER.______

CALVIN COOLIDGE's REVENGE

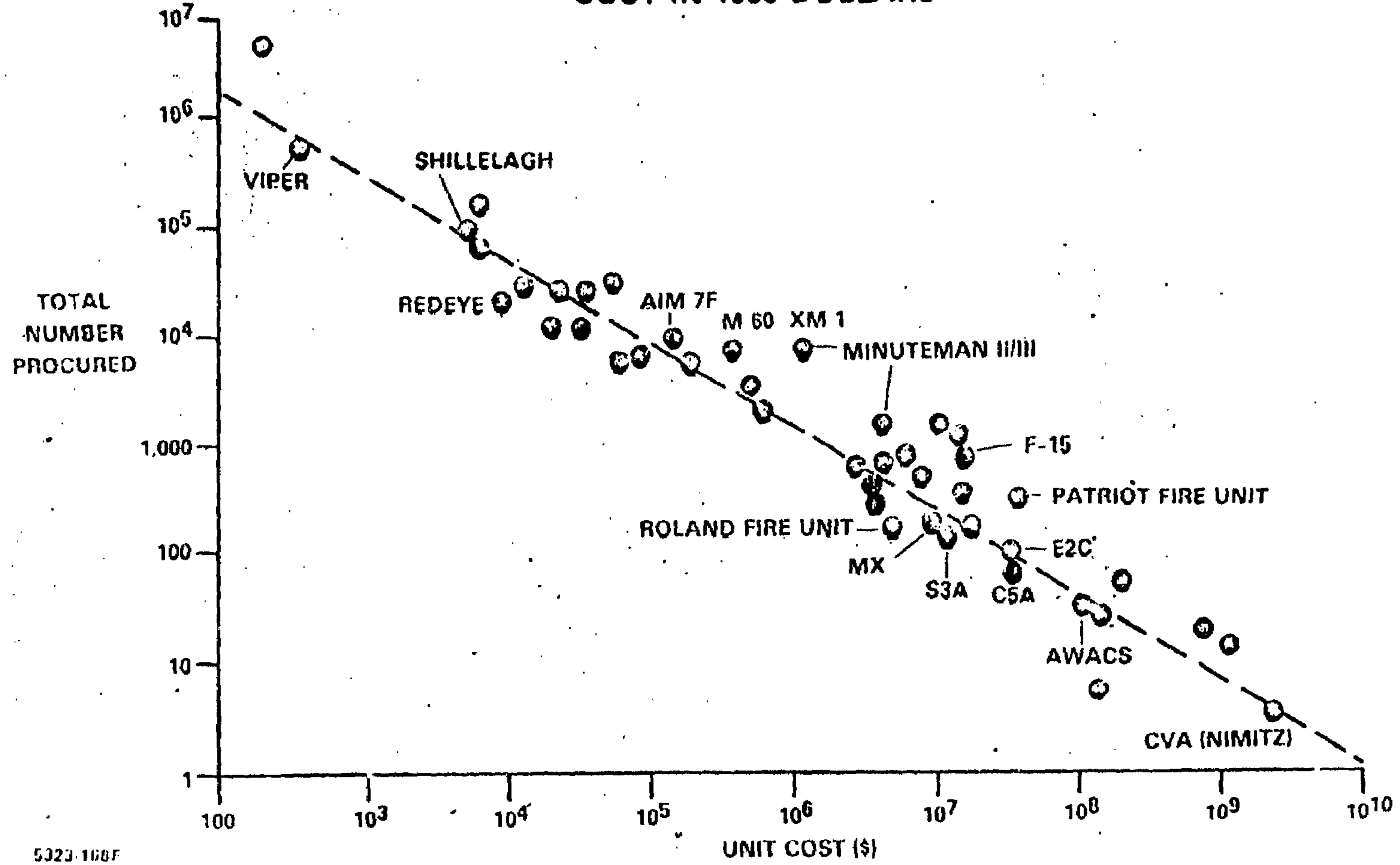


ORGANIZATION OF THE
JOINT CHIEFS OF STAFF
ATE OF ADMINISTRATIVE SERVICES
GRAPHICS BRANCH

8

COST-QUANTITY TRADEOFFS IN MILITARY HARDWARE

COST IN 1980 DOLLARS

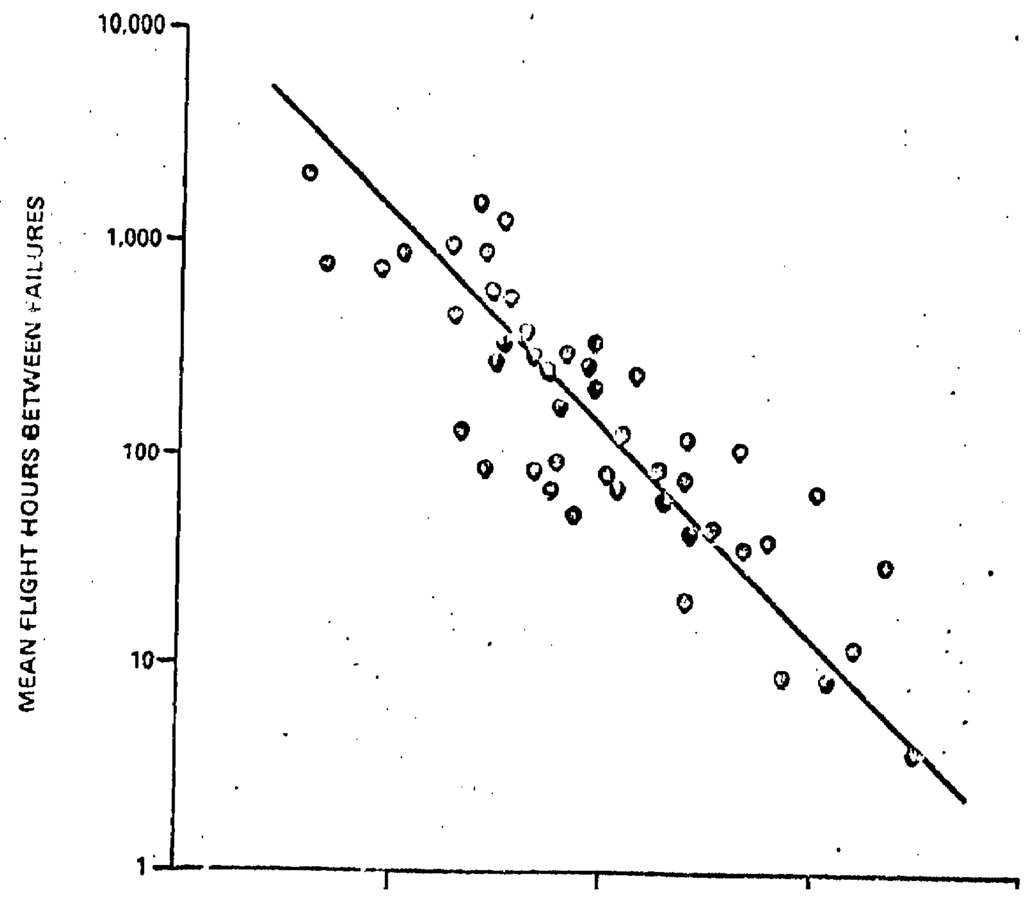


5323-168F

ILE NUMBER

GANIZATION OF THE
INT CHIEFS OF STAFF
OF ADMINISTRATIVE SERVICES
APHICS BRANCH

IMPACT OF INCREASING UNIT COST ON FIELD RELIABILITY 9



GANIZATION OF THE
NT CHIEFS OF STAFF
OF ADMINISTRATIVE SERVICES
APHICS BRANCH

10

WE FIND NO GENERAL EVIDENCE THAT HIGH PERFORMANCE LEADS AUTOMATICALLY TO LOW READINESS

45

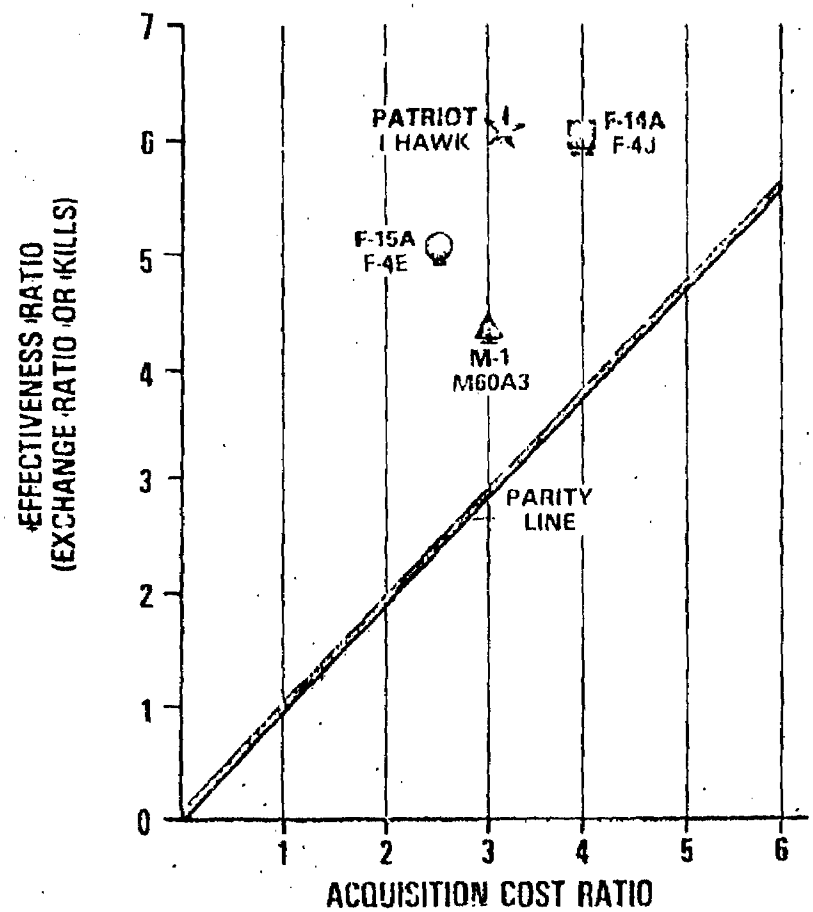
DEFENSE SCIENCE BOARD
SUMMER STUDY PANAL
1981

5323-166H

ORGANIZATION OF THE
JOINT CHIEFS OF STAFF
ORATE OF ADMINISTRATIVE SERVICES
GRAPHICS BRANCH

11

EFFECTIVENESS VS. ACQUISITION COST OF HIGH PERFORMANCE SYSTEMS



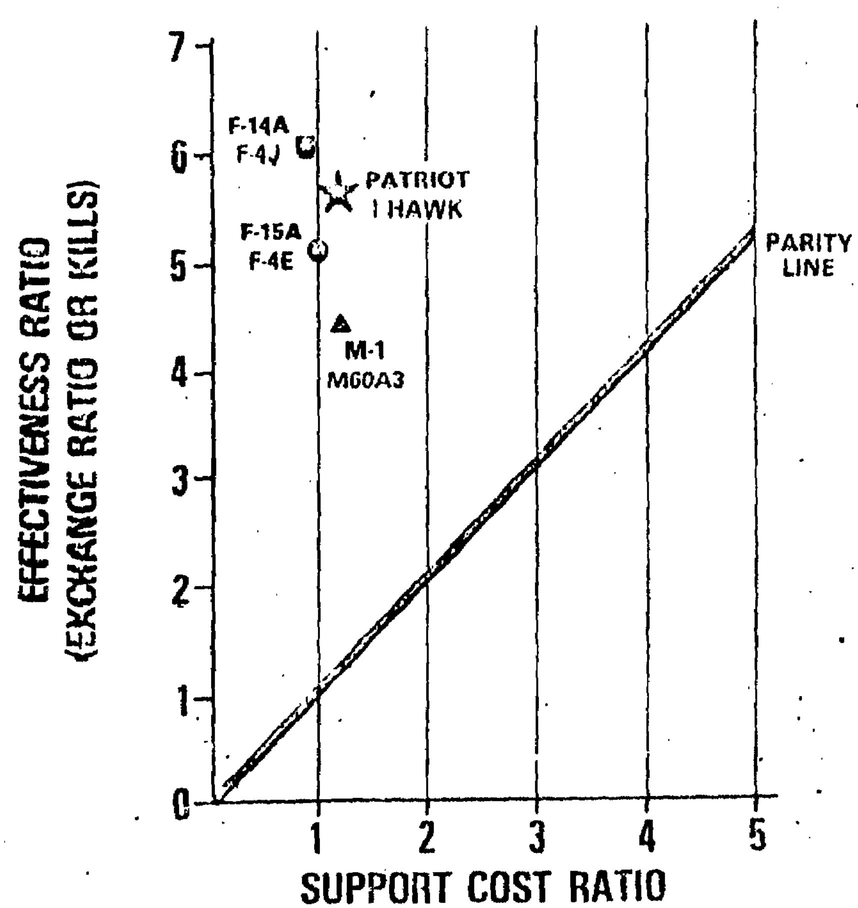
5323-168"1"

DSB '81

ORGANIZATION OF THE
JOINT CHIEFS OF STAFF
DIRECTORATE OF ADMINISTRATIVE SERVICES
GRAPHICS BRANCH

12

EFFECTIVENESS VS. SUPPORT COST OF HIGH PERFORMANCE SYSTEMS



5373-168J

OSD '81

ORGANIZATION OF THE
JOINT CHIEFS OF STAFF
TORATE OF ADMINISTRATIVE SERVICES
GRAPHICS BRANCH

13

THE CASE FOR HIGH TECHNOLOGY WEAPON SYSTEMS

- SOVIETS BUILD MANY OF SAME
- 24 HRS/DAY, MORE "SORTIES" OVERALL
- HUMAN COSTS LOWER
- SUPPORT COSTS LOWER
- LESS STRAIN ON BASING

5323-168L

ORGANIZATION OF THE
JOINT CHIEFS OF STAFF
ATE OF ADMINISTRATIVE SERVICES
GRAPHICS BRANCH

14

TECHNOLOGIES THAT COULD MAKE AN ORDER OF MAGNITUDE DIFFERENCE

	OPPORTUNITY/RISK
1. VERY HIGH-SPEED INTEGRATED CIRCUITS	63/13.5 = 4.7
2. STEALTH	65 5/22 = 3.0
3. ADVANCED SOFTWARE/ALGORITHM DEVELOPMENT	71/29.8 = 2.4
4. MICROPROCESSOR-BASED PERSONAL LEARNING AIDS	41/17 = 2.4
5. FAIL-SOFT/FAULT-TOLERANT ELECTRONICS	55.5/25 = 2.2
6. RAPID SOLIDIFICATION TECHNOLOGY	46/21 = 2.1
7. MACHINE INTELLIGENCE	65.5/31.7 = 2.1
8. SUPERCOMPUTERS	52.5/24.5 = 2.1
9. ADVANCED COMPOSITES	43/21.3 = 2.0

5323-169E

DSB '81

15

TECHNOLOGY SYSTEMS/APPLICATIONS	OPPORTUNITY/RISK NATURE OF IMPACT*	INVESTMENT STATUS
3. ADVANCED SOFTWARE/ ALGORITHM DEVELOPMENT	71/29.2–2.4	$7M
○ CRYPTOGRAPHY ○ TARGET ACQUISITION ○ SOFTWARE MAINTENANCE ○ PERVASIVE ACROSS DOD SYSTEMS	○ PORTABILITY OF SOFTWARE ○ FASTER SOFTWARE DESIGN/TESTING/ MAINTENANCE* ○ LESS EXPENSIVE ALTERNATIVES TO HARDWARE	○ BADLY UNDERFUNDED GIVEN IMPACT POTENTIAL ○ POORLY COORDINATED AND CONCEIVED ○ FRAGMENTED ○ BEHIND STATE OF THE ART

* ORDER OF MAGNITUDE IMPROVEMENT

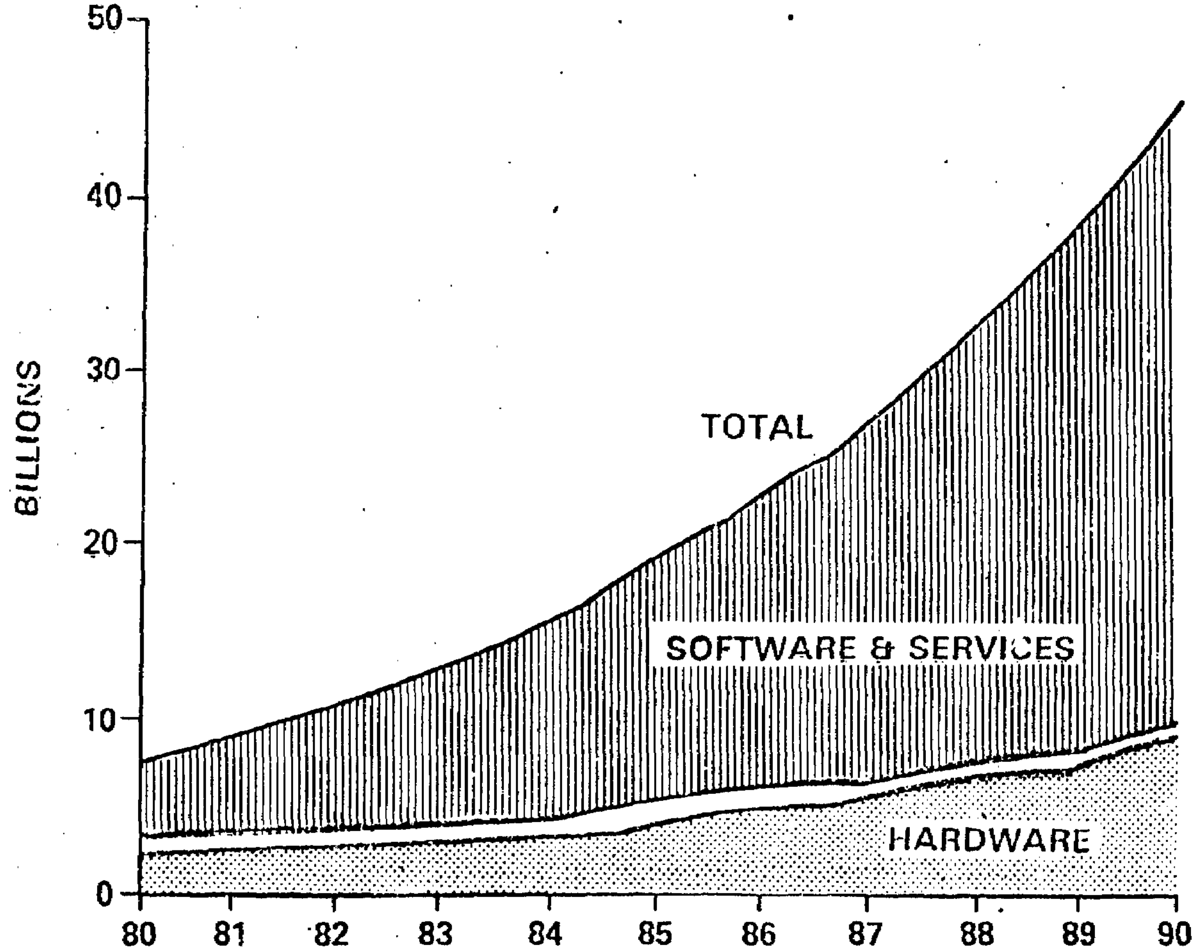
5323-169H

DSB '81

ORGANIZATION OF THE
JOINT CHIEFS OF STAFF
RATE OF ADMINISTRATIVE SERVICES
GRAPHICS BRANCH

16

DEFENSE COMPUTER EXPENDITURE



5323-168U

17

TECHNOLOGY SYSTEMS/APPLICATIONS	OPPORTUNITY/RISK NATURE OF IMPACT*	INVESTMENT STATUS
5. FAIL SOFT/FAULT TOLERANT ("SELF POLICING") ELECTRONICS	55.5/25 = 2.2	\$5M (NO CENTRAL PROGRAM)
○ PERVASIVE IMPACT ON SYSTEM AVAILABILITY	○ ELECTRONIC SYSTEM AVAILABILITY AND RELIABILITY* ○ LIFE CYCLE COSTS* ○ SIMPLIFIED TEST	○ NO COHERENT PROGRAM FOCUS ○ POOR TRANSITION OF PROMISING TECHNOLOGY ALTERNATIVES

* ORDER OF MAGNITUDE IMPROVEMENT

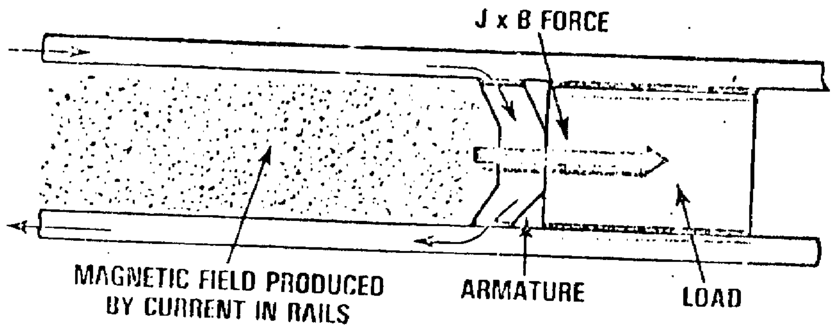
TECHNOLOGY	OPPORTUNITY/RISK	INVESTMENT STATUS
SYSTEMS/APPLICATIONS	NATURE OF IMPACT*	
7. MACHINE INTELLIGENCE	65.5/31.7 = 2.1	$13.2M
○ C^2 AND CRISES MANAGEMENT ○ AUTONOMOUS WEAPONS ○ "INTELLIGENT" DATA BASES	○ PERVASIVE DEPLOYMENT OF COMPUTERS* ○ DISTRIBUTED INFORMATION ○ RECOGNITION AND CORRELATION* ○ NATURAL INTERFACING* ○ TRANSPARENT COMPLEXITY* ○ AUTOMATIC PROGRAMMING	○ PREDOMINATELY DARPA ○ SLOW SERVICE PICKUP ○ APPLICATIONS LAGGING ○ UNDERFUNDED BUT TALENT LIMITED

* ORDER OF MAGNITUDE IMPROVEMENT

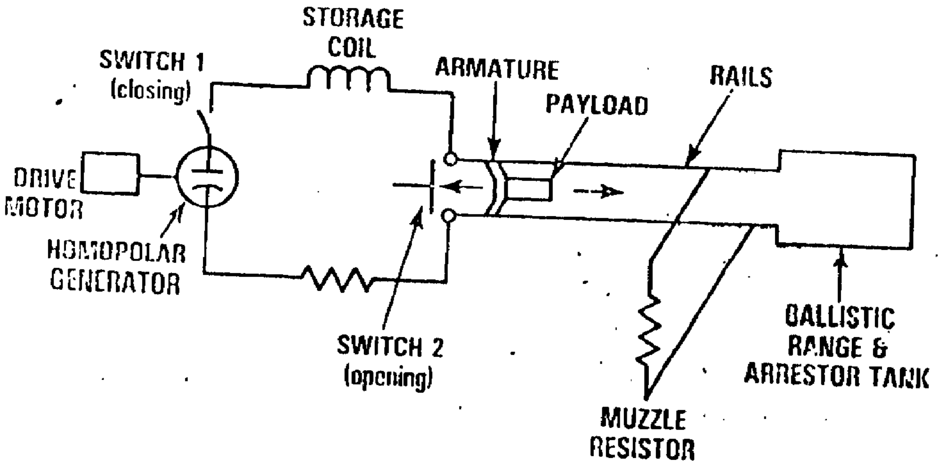
5323-169L

DSB '81

PRINCIPLES OF THE DC PARALLEL RAIL LAUNCHER



OPERATION OF DC EM LAUNCHER



FEATURES OF EM LAUNCHERS

- hypersonic muzzle velocities
- Reduced time-of-flight — improved P_h
- Increased impact effects — improved P_k
- No propellant storage — reduced vulnerability
- Wide range of potential applications
  - Air defense guns
  - Tank cannon
  - RPV launchers
  - Missile launchers

- Inertial energy storage in HPG from prime mover
- Energy compression into inductor during charging by closing switch 1
- Opening switch 2 transfers inductive energy into armature and launches projectile.
- Muzzle resistor dissipates unused stored energy

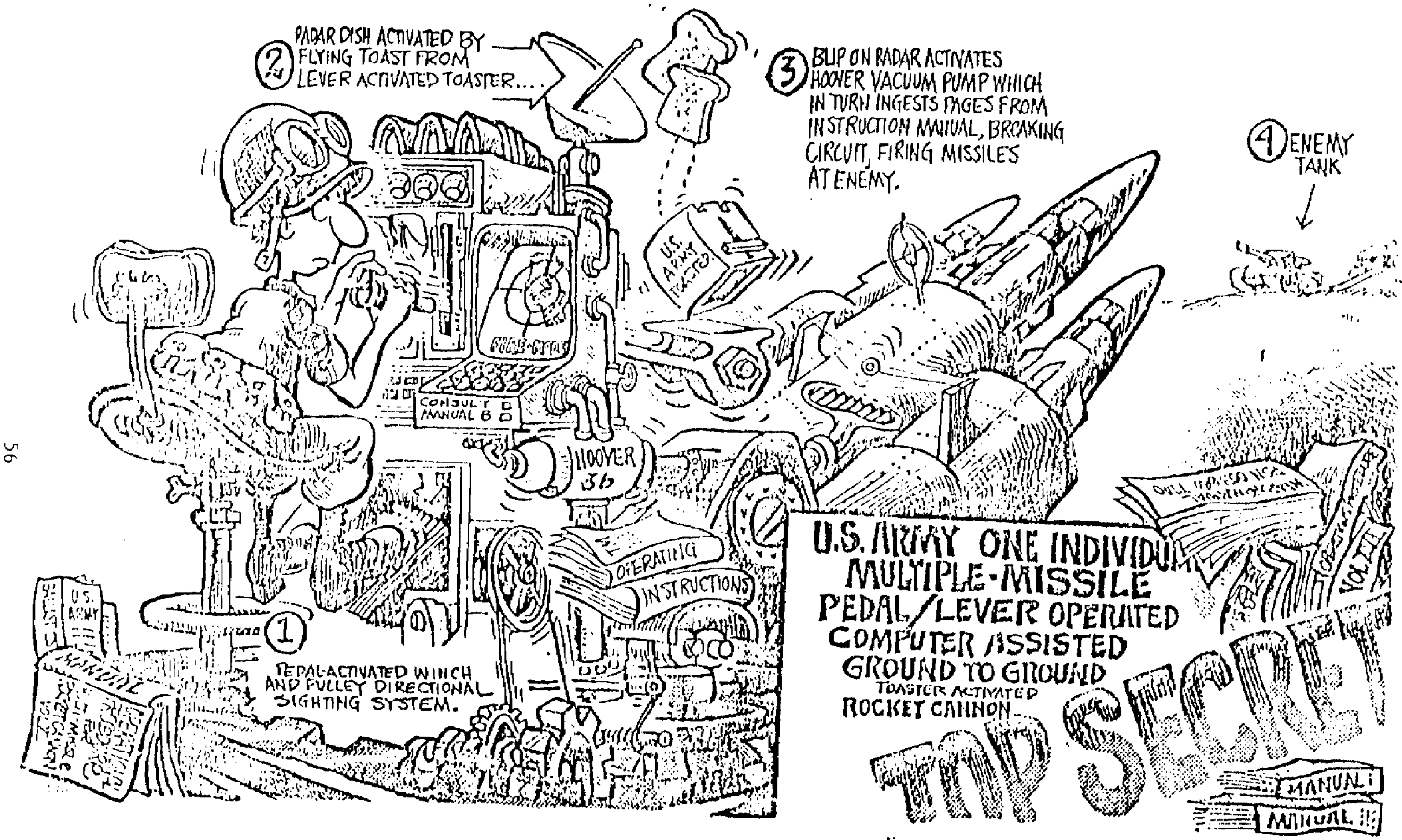
ORGANIZATION OF THE
JOINT CHIEFS OF STAFF
RATE OF ADMINISTRATIVE SERVICES
GRAPHICS BRANCH

FILE NUMBER.

20

8 INCH PROJECTILES

- STORED ON GROUND INDIVIDUALLY
- LIFTED TO TRUCK BED ONE-BY-ONE, BY HAND
- LIFTED TO GROUND ONE-BY-ONE, BY HAND
- LIFTED TO TRUCK BED ONE-BY-ONE, BY HAND
- TRANSFERRED TO HOWITZER ONE-BY-ONE, BY HAND
- LOADED ONE-BY-ONE, BY HAND

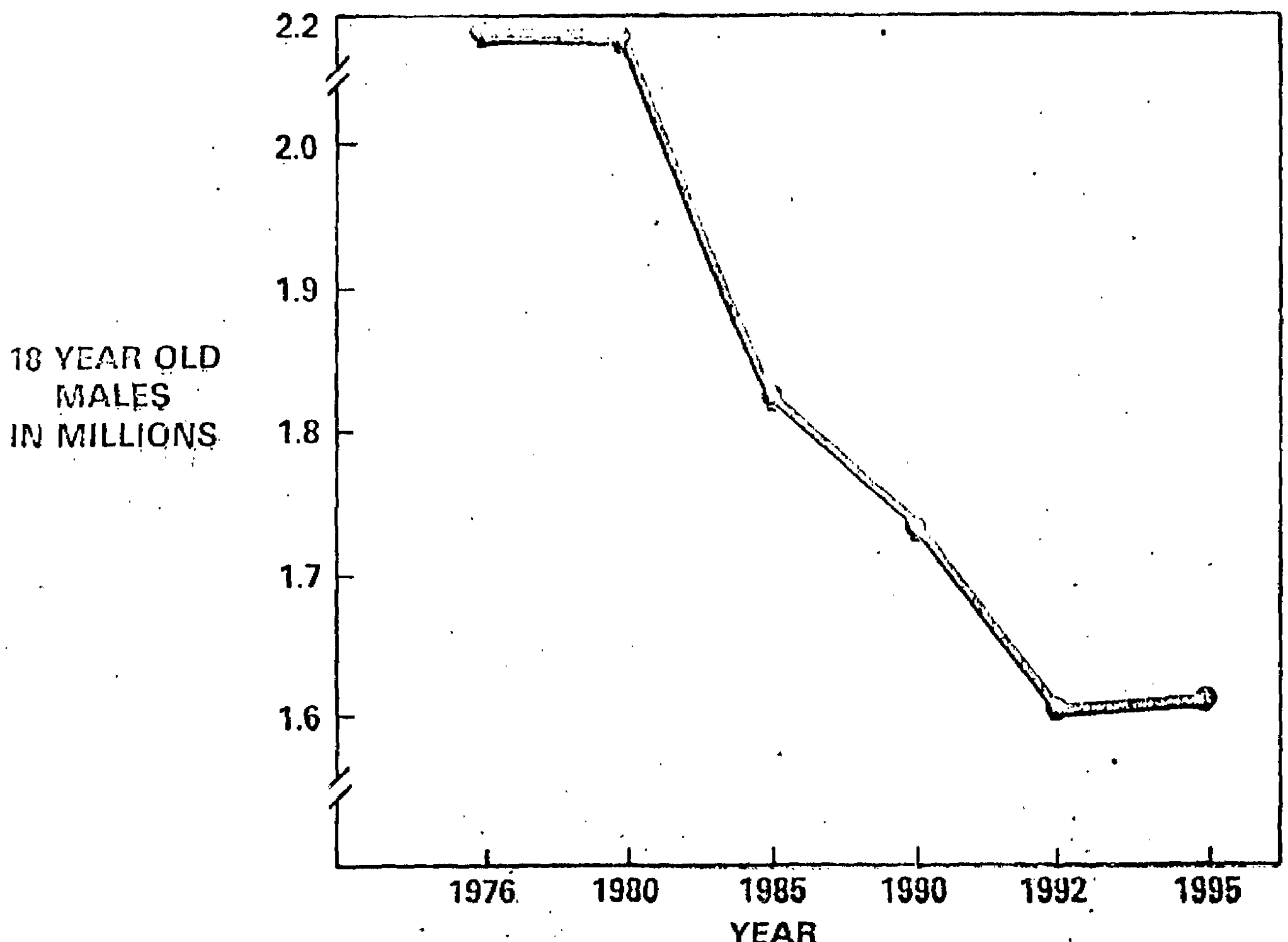


By William T. Coulter

JOINT CHIEFS OF STAFF
DIRECTORATE OF ADMINISTRATIVE SERVICES
GRAPHICS BRANCH

aa

MANPOWER PROJECTIONS, 1980's AND 1990's



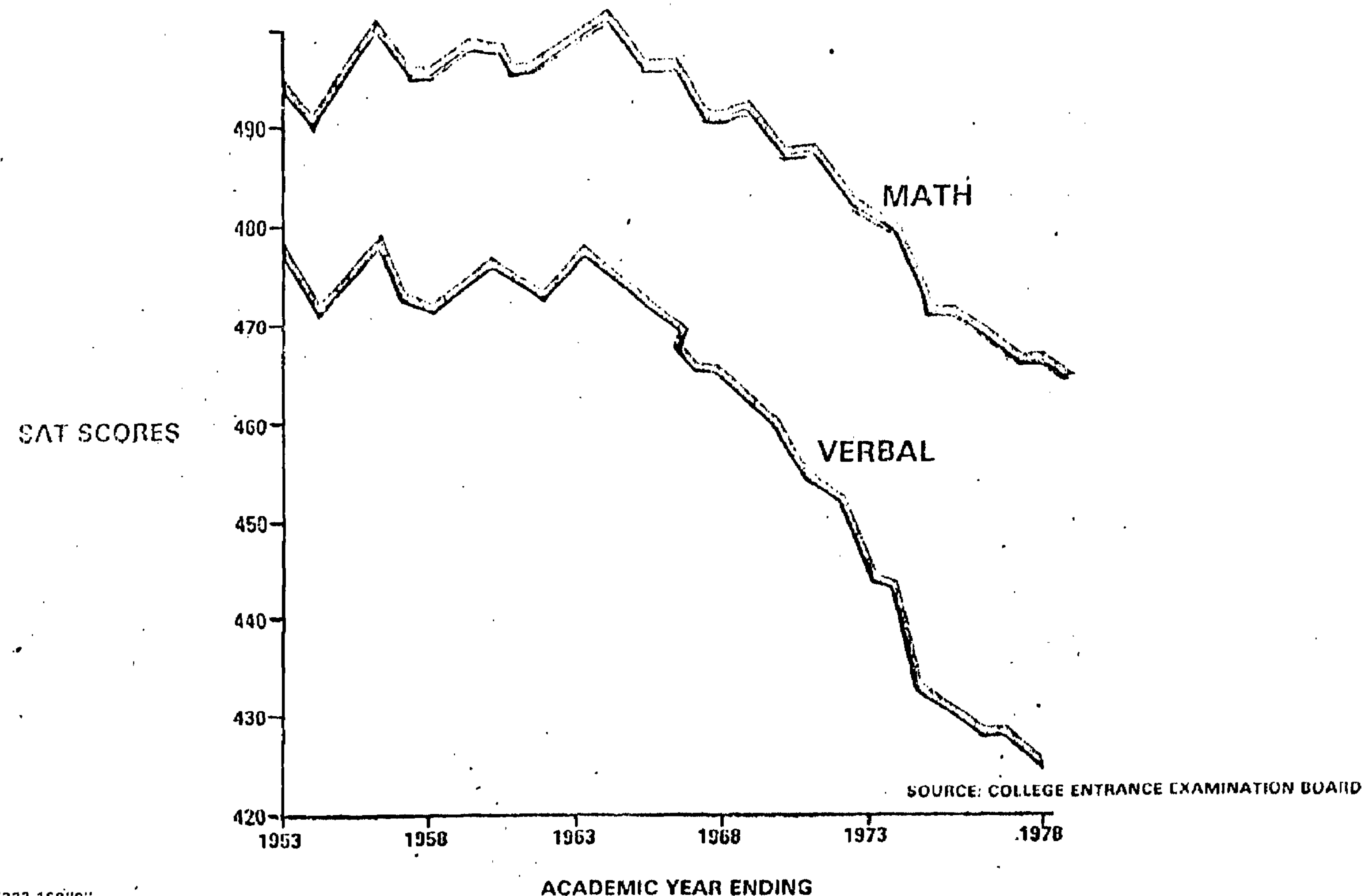
5323-169T

ORGANIZATION OF THE
JOINT CHIEFS OF STAFF
[?]RATE OF ADMINISTRATIVE SERVICES

GRAPHICS BRANCH

aa

SAT SCORES



5323-168"0"

ORGANIZATION OF THE
JOINT CHIEFS OF STAFF
TE OF ADMINISTRATIVE SERVICES

RAPHICS BRANCH

FILE NUMBER: ____

SOVIET VS U.S. SECONDARY SCHOOL GRADUATE

24

	SOV (YEARS)
ALGEBRA	+1.5
GEOMETRY	+8.0
CALCULUS	+1.5
PHYSICS	+4.0
CHEMISTRY	+3.0
BIOLOGY	+3.5
ASTRONOMY	+1.0
MECHANICAL DRAWING	+3.0
WORKSHOP	+8.0

NSF REPORT, 1979

ORGANIZATION OF THE
JOINT CHIEFS OF STAFF
IATE OF ADMINISTRATIVE SERVICES
GRAPHICS BRANCH

FILE NUMBER

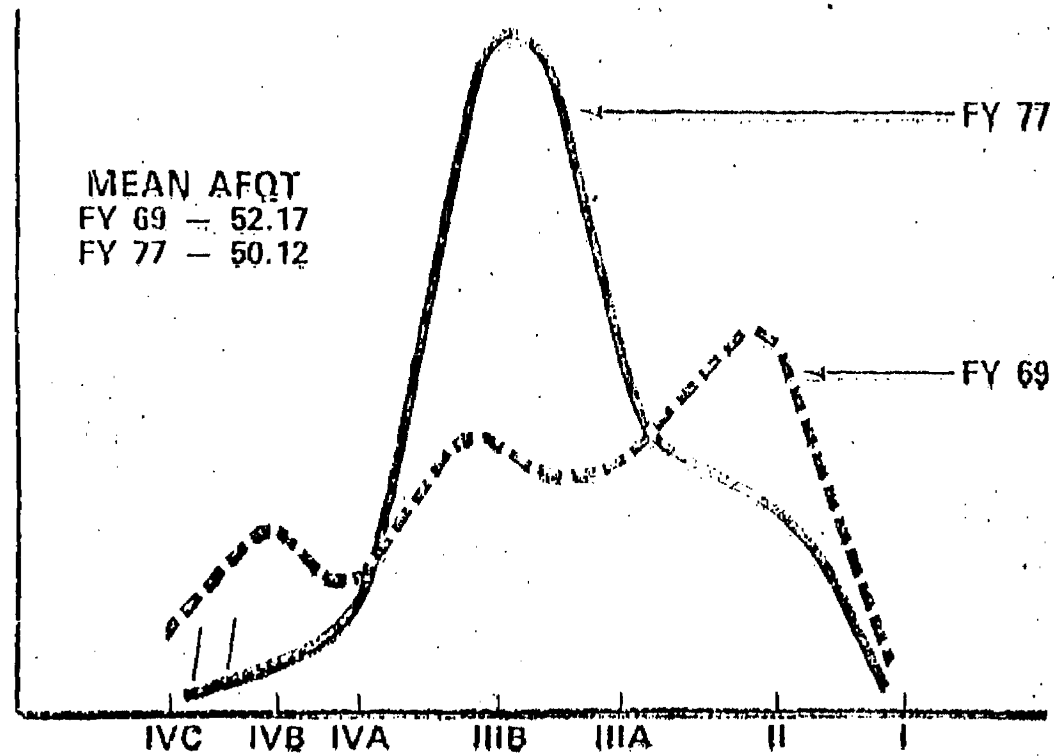
25

NSF, 1979

"THE DISPARITY BETWEEN THE LEVEL OF TRAINING IN SCIENCE AND MATHEMATICS OF AN AVERAGE SOVIET SKILLED WORKER OR MILITARY RECRUIT...AND AN AVERAGE MEMBER OF OUR ALL-VOLUNTEER ARMY IS SO GREAT THAT COMPARISONS ARE MEANINGLESS."

a6

DISTRIBUTION BY MENTAL CATEGORY = MALES

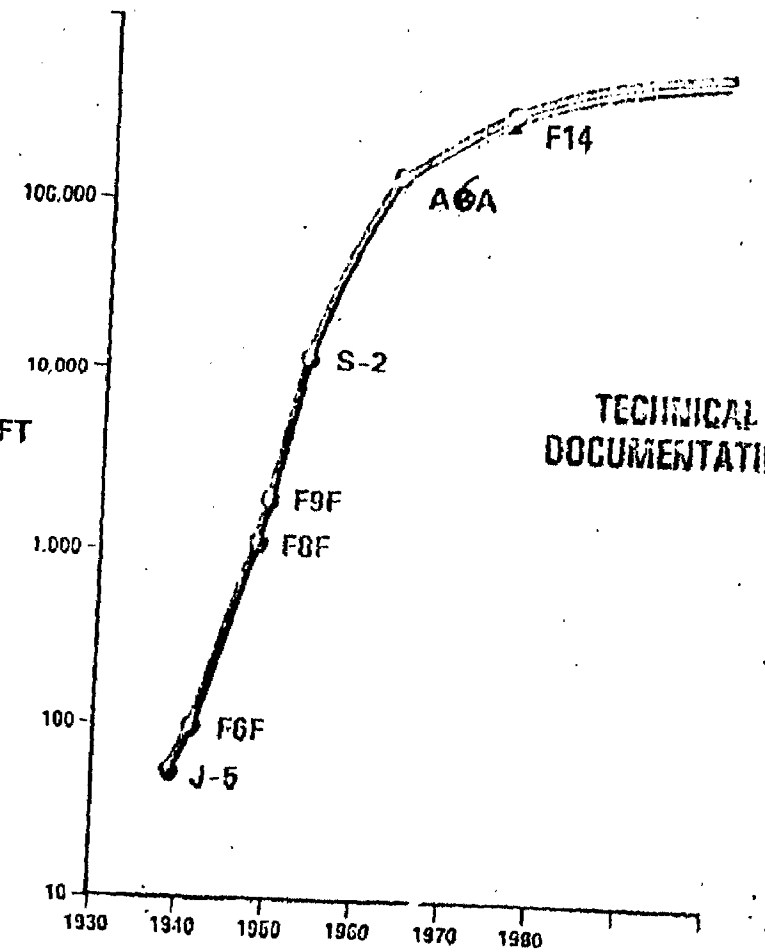


FY 77 ACCESSIONS PEAKED AT MENTAL CATEGORY IIIB REPRESENTING PERCENTILE RANGE OF 31 TO 49 IN MILITARY APTITUDE

ORGANIZATION OF THE
JOINT CHIEFS OF STAFF
IRATE OF ADMINISTRATIVE SERVICES
GRAPHICS BRANCH

27

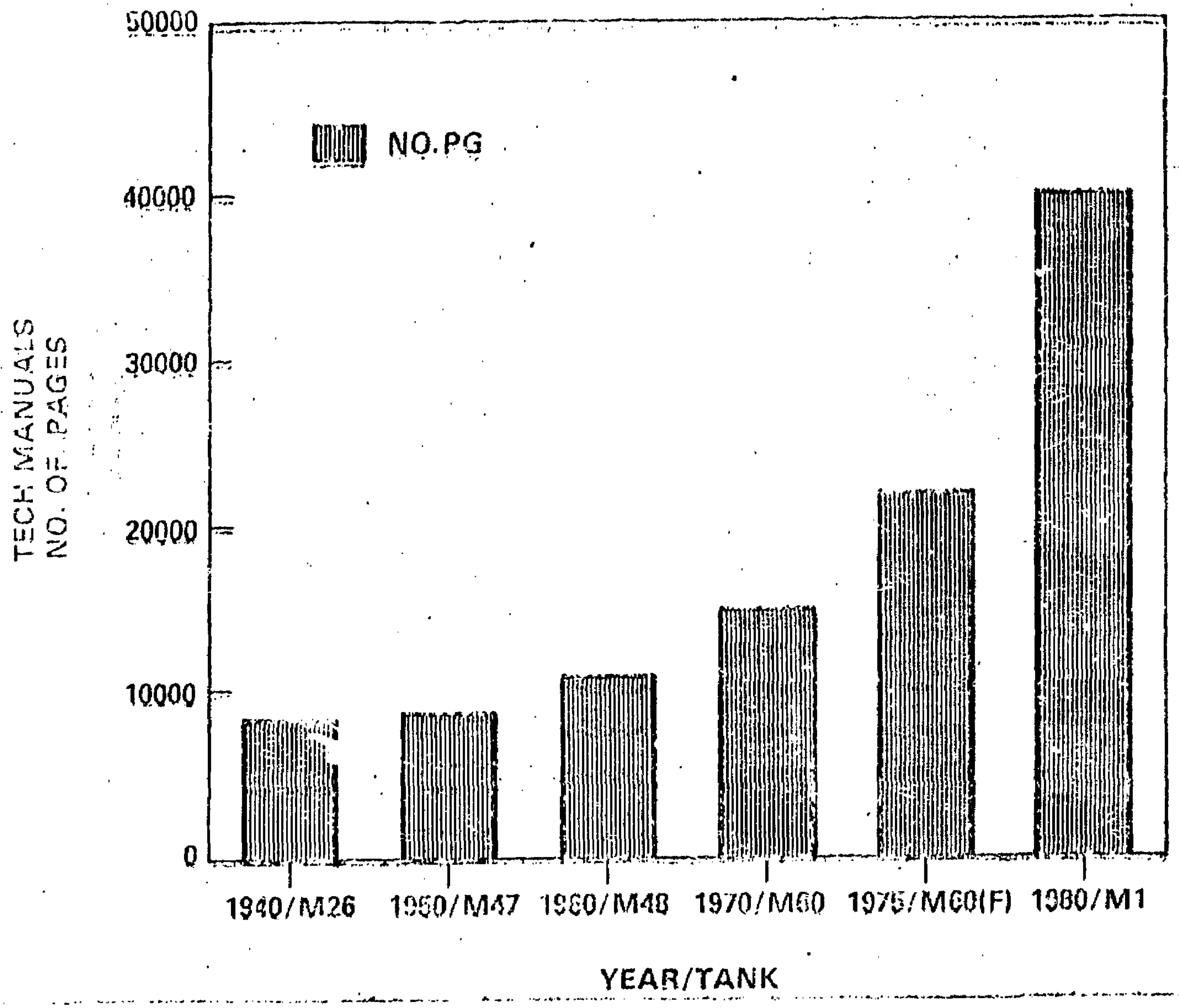
PAGES IN AIRCRAFT MANUALS



TECHNICAL DOCUMENTATION

5323-168M

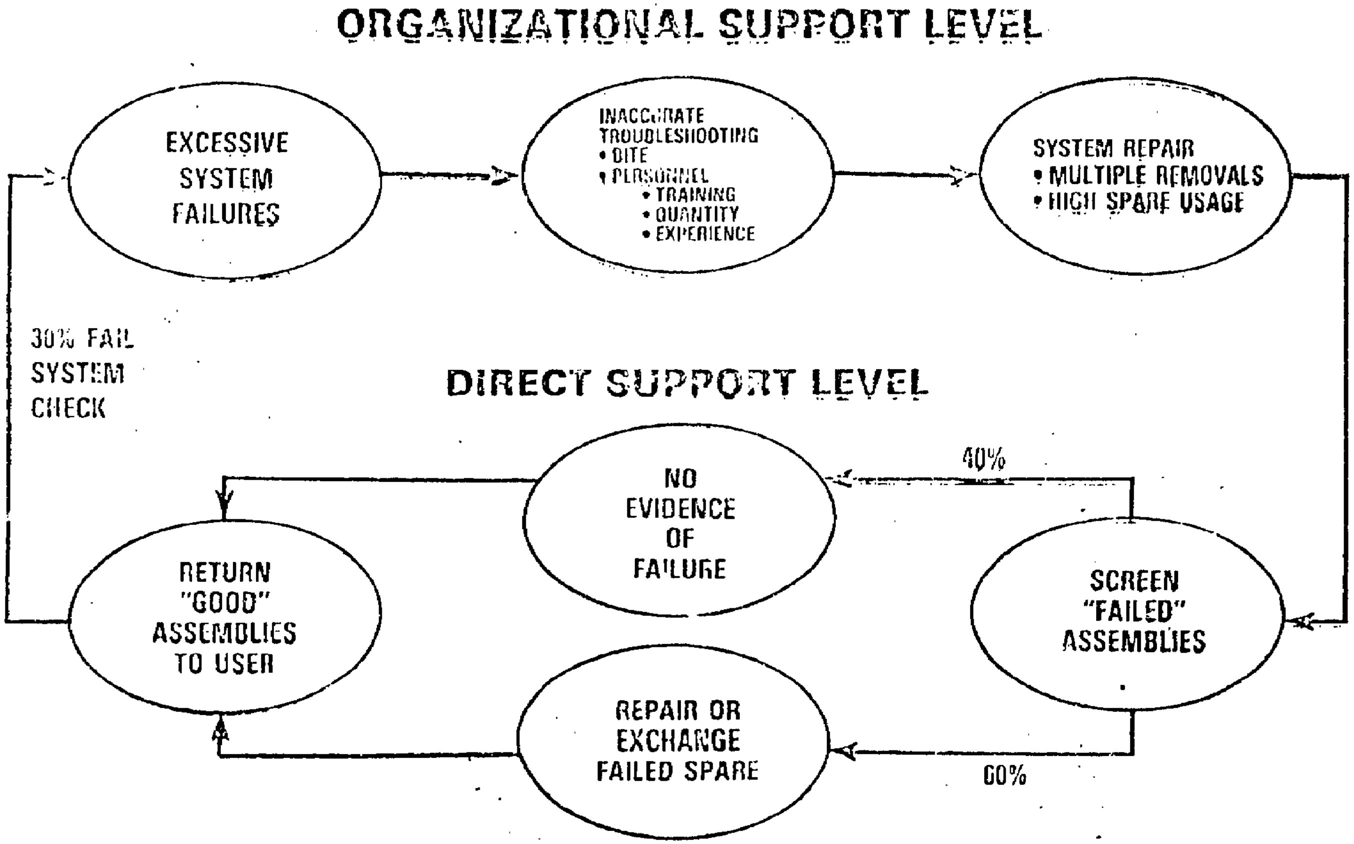
28



NO.PG
TECH MANUALS
NO. OF PAGES
50000
40000
30000
20000
10000
0
1940/M26
1950/M47
1960/M48
1970/M60
1975/M60(F)
1980/M1
YEAR/TANK

29

MANIFESTATIONS OF MAINTENANCE COMPLEXITY (I HAWK EXAMPLE)



5323-168S

IMPACT OF MAINTENANCE COMPLEXITY (I HAWK COMPARISON)

OVERALL CRITERION MEASURES	U.S. ARMY	GERMAN AIR FORCE
NO EVIDENCE-OF-FAILURE RATE	40%	2%
FAILED-SYSTEM-CHECK RATE	30%	UNKNOWN
REPAIR TIME DISTRIBUTION	NOT AVAILABLE	

5323-168R

ORGANIZATION OF THE
JOINT CHIEFS OF STAFF
DIRECTORATE OF ADMINISTRATIVE SERVICES

GRAPHICS BRANCH

31

APPARENT COMPLEXITY
(I HAWK COMPARISON)

INDICATORS	USA-GAF COMPARISON	
MANPOWER AUTHORIZATION BY MOS	SIMILAR	
SKILL LEVELS (GRADE)	USA 2 GRADES LOWER THAN GAF	
SKILL UTILIZATION	SIMILAR DIVERSIONS AND WORKHOURS	
INDIVIDUAL DIFFERENCES	USA	GAF
◦ CIVILIAN EDUCATION ◦ READING GRADE LEVEL	67% HIGH SCHOOL/GED AVG. 8 TO 9	100% SEC + VOCATIONAL AVG. 11 TO 12
ACCESSION POPULATION FOR MAINTENANCE MOS	83% 3-YEAR 16% 4-YEAR	8 YEAR OBLIGORS WITH OVER 3 YEARS OPERATOR EXPERIENCE

6323-100Q

ORGANIZATION OF THE
JOINT CHIEFS OF STAFF
ATE OF ADMINISTRATIVE SERVICES
GRAPHICS BRANCH

32

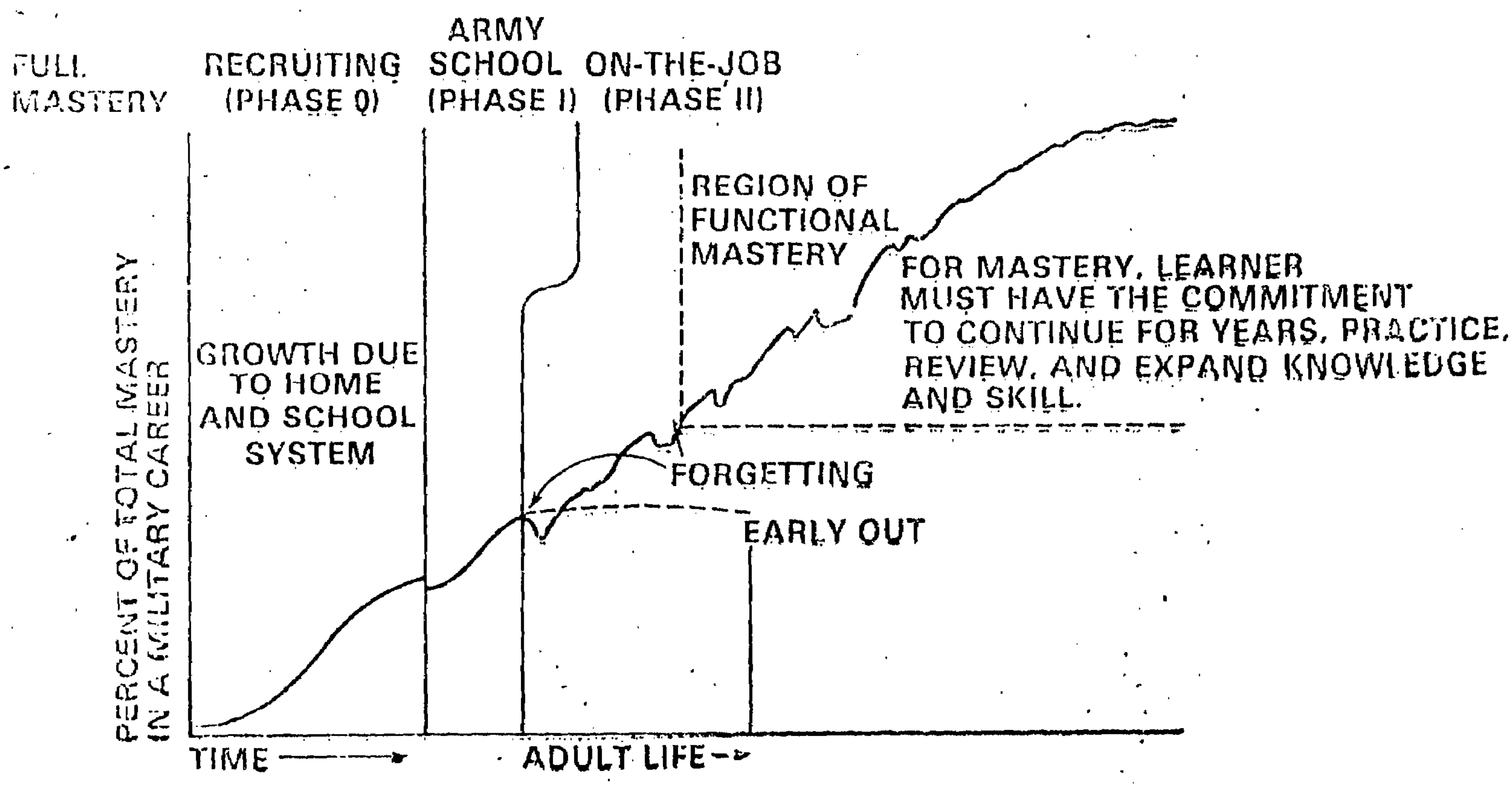
EL SCORE DISTRIBUTION

		PERCENT OF INVENTORY BY EL CATEGORY		
		EL ≤ 99	100 ≤ EL ≤ 109	EL ≥ 110
TOTAL ACTIVE ARMY		49%	21%	30%
I HAWK	ORG	15%	23%	62%
	DS/GS	2%	24%	74%

5323-168P

3

THREE PHASES IN ACHIEVING MASTERY IN A MILITARY CAREER



34

TECHNOLOGY	OPPORTUNITY/RISK	INVESTMENT STATUS
SYSTEMS/APPLICATIONS	NATURE OF IMPACT*	
4. MICROPROCESSOR-BASED PERSONAL LEARNING AIDS	41/17 2.4	$0M
○ TROOP TRAINING • PROCEDURES • MAINTENANCE ○ WEAPON LAYING	○ COMPRESS TRAINING TIME ○ HIGHER SKILL LEVELS ○ TRAINING COST FLEXIBILITY* ○ EXTENDS MANPOWER POOL ○ AIDS RETENTION	○ AN OPPORTUNITY DESERVING FUNDING $4M NEEDED ○ DEVELOPMENT OF TRAINING SOFTWARE FOR PERSONAL COMPUTERS ○ AN EXPERIMENT PROGRAM OF SUFFICIENT SIZE TO JUDGE EFFECTIVENESS

* ORDER OF MAGNITUDE IMPROVEMENT

5323-169G

DSB '81

PAPER PRESENTATIONS

Adams, Jerome, MAJ, United States Military Academy, West Point, New York & Hicks, Jack M., US Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia. (Tues. P.M.)

Performance of Male and Female Cadets during Cadet Field Training

All female cadets (N = 86) and a random sample of male cadets (N = 49) in the Class of 1982 were compared in terms of the performance ratings they received at Cadet Field Training (CFT). Statistically significant results from multivariate analysis of variance indicated that male cadets were evaluated more favorably than were female cadets. However, the magnitude of these gender differences was small. Other effects indicated that: (1) regular Army officers rated cadets less favorably than did more senior cadets; (2) male raters did not differ from female raters in how they rated cadets at CFT; (3) cadets in the squad member role were rated lower than cadets in leadership or administrative roles; and (4) the difference between ratings of male and female cadets was less in the squad member role than in either leader or administrator roles. A sex role stereotype interpretation was suggested for the gender effect shown with the CFT rating form. Possible directions for change in the rating form were proposed. Some cautions regarding source and context effects were suggested for Academy personnel taking action on the basis of CFT performance ratings.

Leadership Performance Appraisal Ratings
During Cadet Field Training

Jerome Adams
Associate Professor
Science Research Lab, USMA
West Point, NY 10996

Jack M. Hicks
Senior Research Psychologist
Army Research Institute
Alexandria, VA 22333

This paper describes the results of analyses comparing the performance appraisal ratings given to male and female cadets during various phases of Cadet Field Training (CFT). CFT takes place at Camp Buckner, a special military field training facility located on the grounds of the United States Military Academy. CFT is scheduled for a seven week period during the summer following completion of the freshman year at the Academy. During this period, the cadets are introduced to an orientation into the combat arms and combat support branches of the Army, e.g. armor, field artillery, infantry, engineering, signal, air defense artillery, etc. Consistent with the physically demanding nature of being in the combat and combat support branches, the CFT training regimen is very demanding in terms of the level of physical effort and endurance involved. Long marches, daily runs and calisthenics, manual manipulation of heavy construction materials, long hours in hot armored vehicles, eating meals in the field including field rations, and too little sleep are just some of the physical rigors which characterized CFT at the time of this study in 1979.

One important goal of CFT is to develop the leadership skills of the participating cadets. In line with this goal, cadets are assigned to temporary leadership during the course of training. For example, a cadet may serve as squad leader or as a section chief or platoon sergeant for a temporary period in training and then, at the end of the same day, the cadet would return to the role of squad member. Thus, there is a continuous rotation of leadership roles.

*The research reported here was supported by grant MDA 903-78-G02 from the U.S. Army Research Institute for the Social and Behavioral Sciences (Jerome Adams, Principal Investigator).

**This paper represents the views of the authors and not the official position of the U.S. Military Academy, the U.S. Army Research Institute, the U.S. Army, or any other governmental agency unless so designated by authorized documents.

***We thank Debra Instone and Robert Rice for their assistance on an earlier report and for their assistance in data analysis.

PRECEDING PAGE BLANK-NOT FILMED

At several points during the course of training, the performance of cadets at CFT is formally evaluated by either upperclass West Point cadets, or by the regular Army officers that supervise cadet training (i.e., Tactical Officers). A cadet may be in a squad member, a leader, or occupy a supervisory staff role at the time of these evaluations. In this context, typical supervisory positions include transportation sergeant, supply sergeant, training sergeant, etc. Such positions carry considerable responsibility, but are outside of the direct line of command and have limited authority when compared to line leader roles such as assistant squad leader, squad leader, platoon leader, first sergeant, etc.

Their performance appraisal ratings, developed and regularly collected under institutional authority, served as the data analyzed for the present report. The major objective of these analyses was to compare the ratings given to male and female cadets as they performed temporary military training exercises in any of these three roles (i.e., member of squad, formal leader, or supervisor).

METHOD

Subjects

Performance appraisal ratings were obtained for all female cadets (N = 86) and a random sample of male cadets (N = 49) enrolled in the Class of 1982 at the United States Military Academy during the summer of 1979. All these cadets were participants in Cadet Field Training, (CFT).

Performance Measures

An 11 item rating form, developed and administered by the Academy, was used to evaluate the performance of cadets as they performed various duties associated with CFT. The forms used by Tactical Officers (TAC) and more senior cadets holding roles in the Cadet Chain of Command (COC) included the same 11 items. Each of the 11 dimensions used in the rating form is described below. (A copy of the form is included as Appendix A).

(1) Sense of responsibility and reliability in the execution of assigned tasks.

(2) Ability and willingness to work in harmony with others.

(3) Ability to grasp a situation, think clearly and develop logical conclusions.

(4) Ability in organizing and directing the efforts of others.

(5) Capacity for increased responsibility; ability to learn rapidly.

(6) Motivates others by his keen interest and personal participation.

(7) Initiative, forcefulness and aggression.

(8) Ability to bear up under pressure; the will to persevere in the face of obstacles.

(9) Ability to adjust to new or changing situations and stresses.

(10) Renders faithful and willing support to superiors and followers.

(11) Overall performance of duties.

The first ten dimensions are rated on a five point scale and the eleventh one uses a ten point scale.

Unit of Analysis

Each cadet was rated by one to five evaluators. Hence, the total number of evaluations was 487 (205 for male cadets 282 for female cadets). In the analyses to follow, the evaluation rather than the individual cadet serves as the unit of analysis (i.e., $N = 487$, not 135).

RESULTS

Intercorrelations

Table 1 presents the intercorrelations among the 11 questions making up the rating scale. These intercorrelations are quite high, suggesting that there is relatively little discrimination in the way raters respond to these 11 different items. Until recently, regular Army officer evaluations were subject to a highly skewed (inflated) performance score. This general leniency error appears to generalize by cadets who typically emulate officer role model behaviors.

Analysis of Variance

A multivariate analysis of variance was conducted. In this 2 x 3 x 11 analysis, Cadet Gender (Male-Female) and Cadet Role (Member of Squad, Leader, Supervisor) served as the independent variables, and the 11 performance ratings served as the dependent variables.

The MANOVA yielded two significant main effects and a marginal interaction effect: Cadet Gender ($F = 4.58$, $df = 22,838$, $p < .00001$); Cadet Role ($F = 2.72$; $df = 22,888$; $p < .00003$); Cadet Gender x Cadet Role ($F = 1.37$; $df = 22,888$; $p < .12$). To interpret the meaning of these significant and marginally significant effects, follow-up univariate analyses were performed.

Because of space limitations, I will only describe briefly the univariate results related to each of the three effects tested in these analyses of the 11 ratings. The complete tabular results are available in the full manuscript.

Main effect for Cadet Gender. Seven of the 11 items yielded significant main effects for Cadet Gender in the univariate analyses of variance. The direction of these effects consistently showed males being rated more favorably than females in every instance of a significant difference. Specifically, male cadets, relative to female cadets, were rated as:

- being more reliable and responsible in executing assigned tasks

- having greater ability to grasp a situation, think clearly, and develop logical conclusions

- having greater ability to organize and coordinate the efforts of others

- having greater capacity for increased responsibility

- more motivating to others through keen interest and personal participation

- having greater initiative, forcefulness, and aggressiveness

- better in overall performance of duties

On only the four following items did the difference in ratings of male and female cadets fail to achieve conventional levels of statistical significance:

- ability and willingness to work in harmony with others

- ability to bear up under pressure; the will to persevere in the face of obstacles

- ability to adjust to new or changing situations or stresses

- renders faithful and willing support to superiors and subordinates

However, for all four of these items the direction of the means was the same as that found in the significant effects, i.e., males were rated higher than females.

Main effects for Cadet Role. The significant multivariate effect for Cadet Role is reflected by significant univariate effects for 10 of the 11 rating dimensions. Cadets in the role of squad member were evaluated least favorably and those occupying an administrative role were evaluated most favorably. Cadets in positions of formal leadership at the time they were rated were the recipients of intermediate evaluations, falling between the other two roles. For some dimensions the mean ratings of cadets in leader roles fell closer to the mean for cadets in squad member roles. While for other dimensions the means for leaders were closer to the means received by cadets holding administrative roles.

Interaction effect of Cadet Gender x Cadet Role. In the univariate analyses of variance, six of the 11 dimensions yielded significant interactions between Cadet Gender and Cadet Role. The means associated with these effects are presented in Table 2 along with the results of Newman-Kuels multiple comparison tests (Winer, 1971, pp. 210-219). Those means sharing the same subscript in Table 2 are not significantly different from each other. Each Newman-Kuels test is limited to a single dependent variable. Thus, in examining the subscripts associated with the means for a significant interaction, one must consider each row of the table separately.

The results of these 2 x 3 interactions are complex. However, it is possible to offer some general comments for these six significant interactions. The smallest difference between the ratings of male and female cadets was found when cadets were members of squad. None of the six dimensions yielding significant interactions showed a significant difference between ratings of male and female cadets while in the role of member of squad. However, for five of the six variables, the difference between ratings of male and female cadets was significant for cadets holding formal leadership roles. In formal leader roles as this type, male leaders, relative to female leaders were rated as:

- having greater capacity for increased responsibility

- having greater ability to organize and coordinate the efforts of others

- having greater initiative, forcefulness, and aggressiveness

- having greater ability to adjust to new or changing situations and stresses

- better in overall performance of duties

Regarding the third role, administrator, the pattern of male-female differences was somewhat mixed. Three variables showed no significant differences between males and females in this role (capacity for increased responsibility; adjusting to new situations; and support to superiors and subordinates). The other three variables yielded significantly higher ratings for male cadets in administrative roles than for female cadets in these roles (organizing and directing efforts of others; initiative, forcefulness, and aggression; overall performance). In sum, these interactions show the little discrepancy between ratings of male and female cadets in follower roles, but higher ratings for male cadets in formal leadership roles, and to a lesser extent higher ratings for males than females in administrative roles.

DISCUSSION

The most noteworthy result of the present study is the consistent difference in the ratings of male and female cadets. Males were rated significantly higher than females on seven of the 11 rating dimensions. Interaction effects indicated that for many of these dimensions, this gender effect was smallest when cadets were in a squad member role and greatest when they held positions of formal leadership or administrative responsibilities. An examination of the correlations between cadet gender and the 11 rating dimensions presented in Table 1 shows that the magnitude of the gender effects in the present study are small in magnitude. The largest correlation involving the cadet gender variable is only .22, indicating less than 5% shared variance (with item 4, organize and direct others).

Many different interpretations could be offered for these small, but statistically significant, effects showing that males are rated more favorably than females in carrying out their duties as CFT. One interpretation of these findings can be offered in terms of sex role stereotypes. Deaux (1976) indicates that females are expected to be compassionate, emotional, warm, quiet, gentle, passive and tactful while males are expected to be analytic, aggressive, independent, assertive, confident, and direct. Schein (1973) demonstrated the relevance of sex role stereotypes for concerns with leadership and managerial roles. She compared the stereotype of males and females with the stereotype of effective managers. Her finding was that "male" and "effective manager" share a large number of common stereotypic attributes while "female" and "effective manager" share few attributes. The implication of Schein's study is that masculinity may be perceived as effectiveness in leadership roles while femininity may be perceived as being less effective.

The Cadet Role effects may reflect nothing more than the behavioral repertoire appropriate for the different roles. It may be very difficult to display some of the behaviors and traits included in the rating scale while in the squad member role. On the other hand, there are abundant opportunities to engage in such behaviors while in leadership or administrative roles. No doubt alternative interpretations could be offered for the phenomena discussed above. While there remains ambiguity in the theoretical meaning of Cadet Role and Rater Role effects, the practical implications of these results are clear. Namely, when acting on CFT performance ratings, Academy personnel have begun to attend to both the source and the context of such ratings. The present results show that it is unjustified to assume comparable meaning for ratings from officers and from cadets, or for cadets occupying different roles in the CFT program. Before taking any administrative actions on the basis of such ratings, Academy officials have come to recognize that it is essential to judge the ratings in light of the source and role from which the ratings are generated.

References

Adams, J., Rice, R.W. and Instone, D., The 1979 Summer Leadership Study: A comparison of male and female leaders (Tech. Rep. 80-2). Army Research Institute for Social and Behavioral Sciences, Grant MDA 903-78-G02, July 1980.

Cohen, E., and Buens, P. SPSS-Manova-Multivariate Analysis of Variance and Covariance. Evanston, Ill.: Vogel Computing Center, Northwestern University, May 1976.

Deaux, K. The Behavior of Women and Men. Monterey, Calif.: Brooks/Cole, 1976.

Rice, R.W., Bender, L.R., and Vitters, A.G. Leader sex, follower attitudes toward women and leadership effectiveness: A laboratory experiment. Organizational Behavior and Human Performance, 1980, 25, 46-78.

Rice, R.W., Yoder, J.D., Adams, J., Priest, R.F., and Prince, H.T. II. Correlates of leadership ratings for male and female military cadets. Paper presented at the convention of the Eastern Psychological Association, Hartford, Conn., April 1980.

Schein, V.E. The relationship between sex role stereotypes and requisite management characteristics. Journal of Applied Psychology, 1973, 57, 95-100.

Winer, B.J. Statistical Principles in Experimental Design. New York: McGraw-Hill, 1971.

Table 1

Intercorrelations Among Rating Criteria and Cadet Gender

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1. Responsibility	--											
2. Work in harmony	.55	--										
3. Think clearly	.55	.48	--									
4. Organize and direct	.54	.41	.58	--								
5. Increased responsibility	.59	.48	.58	.52	--							
6. Motivate others	.52	.58	.47	.50	.51	--						
7. Initiative	.51	.36	.50	.57	.48	.53	--					
8. Bear up	.48	.41	.49	.43	.47	.45	.50	--				
9. Adjust	.53	.43	.55	.45	.50	.43	.50	.66	--			
10. Render support	.60	.61	.44	.41	.48	.52	.35	.43	.48	--		
11. Overall performance	.74	.61	.62	.64	.54	.65	.64	.63	.63	.62	--	
12. Cadet gender (M = 1, F = 2)	-.11	-.06	-.15	-.22	-.10	-.10	-.18	-.06	-.08	.00	-.17	--
$\bar{X}$	3.79	3.85	3.52	3.48	3.74	3.45	3.32	3.63	3.58	4.04	7.38	1.58
SD	.84	.81	.70	.80	.69	.86	.87	.83	.73	.82	1.57	.49

Note: N = 479 - 481, $r \geq .10$ is significant with $p < .05$ and $r >$ .13 is significant ...

Table 2

Cadet Gender x Cadet Role Interactions

			Mean Ratings					
			Member of Squad		Leader		Administrator	
Variable	F	df	Male	Female	Male	Female	Male	Female
1. Responsibility	3.01*	2,454	3.76_a	3.67_a	4.26_b	3.72_a	4.56_b	4.20_b
2. Work in harmony	1.00	2,454	3.81	3.76	4.21	3.90	4.00	4.05
3. Think clearly	.82	2,454	3.56	3.36	3.94	3.54	4.06	3.85
4. Organize and direct	4.24*	2,454	3.53_a	3.27_a	4.15_b	3.43_a	4.31_b	3.60_a
5. Increased responsibility	3.40*	2,454	3.73_a	3.63_a	4.18_b	3.72_a	4.00_b	4.15_b
6. Motivate others	1.83	2,454	3.46	3.39	3.77	3.43	3.94	3.40
7. Initiative	3.97*	2,454	3.38_a	3.16_a	4.00_b	3.24_a	3.93_b	3.40_a
8. Bear up	2.15	2,454	3.56	3.53	4.09	3.65	4.00	3.85
9. Adjust	3.70*	2,454	3.54_a	3.49_a	4.00_b	3.48_a	4.12_b	4.00_b
10. Render support	3.91*	2,454	3.93_a	4.03_{ab}	4.35_{ab}	3.91_a	4.44_b	4.45_b
11. Overall performance	4.91**	2,454	7.44_a	7.07_a	8.71_b	7.26_a	8.75_b	7.90_a

*$p < .05$ **$p < .01$

Cell means sharing the same subscript are not significantly different from one another at $p < .05$ using Newman-Kuels. Such differences are reported only for variables yielding significant interaction effects.

APPENDIX A

TACTICAL OFFICER BUCKNER PERFORMANCE REPORT

SECTION A

NAME - LAST, FIRST, MI	BUCKNER CO	REGULAR CO

CHAIN OF COMMAND POSITIONS	1ST DETAIL	2D DETAIL

SECTION B - To be completed by Tactical Officer concerned (Enter comments to support ratings)

1. Sense of responsibility and reliability in the executing of assigned tasks.

☐ SUP ☐ AE ☐ EXP ☐ BE ☐ UNSAT

2. Ability and willingness to work in harmony with others.

☐ SUP ☐ AE ☐ EXP ☐ BE ☐ UNSAT

3. Ability to grasp a situation, think clearly and develop logical conclusions.

☐ SUP ☐ AE ☐ EXP ☐ BE ☐ UNSAT

4. Ability in organizing and directing the efforts of others.

☐ SUP ☐ AE ☐ EXP ☐ BE ☐ UNSAT

5. Capacity for increased responsibility, ability to learn rapidly.

☐ SUP ☐ AE ☐ EXP ☐ BE ☐ UNSAT

6. Motivates others by his keen interest and personal participation.

☐ SUP ☐ AE ☐ EXP ☐ BE ☐ UNSAT

7. Initiative, forcefulness and aggressiveness.

☐ SUP ☐ AE ☐ EXP ☐ BE ☐ UNSAT

8. Ability to bear up under pressure; the will to persevere in the face of obstacles.

☐ SUP ☐ AE ☐ EXP ☐ BE ☐ UNSAT

9. Ability to adjust to new or changing situations and stresses.

☐ SUP ☐ AE ☐ EXP ☐ BE ☐ UNSAT

10. Renders faithful and willing support to superiors and subordinates.

☐ SUP ☐ AE ☐ EXP ☐ BE ☐ UNSAT

11. Overall performance of duty. (Comment in any area where cadet has received a rating of BE or UNSAT. Include comments on special abilities, self-improvement efforts and general attitude.)

☐ SUPERIOR ☐ ABOVE EXPECTED ☐ EXPECTED ☐ BELOW EXPECTED ☐ UNSAT

DATE	RANK AND NAME	SIGNATURE

Adams, Jerome, MAJ, Richards, John, CPT & Fullerton, Terry, CPT, United States Military Academy, West Point, New York. (Wed. A.M.)

Relationship between Attitudes and Leadership Style: A Policy Capturing Approach

The present study sought to extend understanding of the relational qualities of leadership style using policy capturing. The approach required analysis of the styles of leadership used by battalion commanders in dealing with problem soldiers. A questionnaire was administered to battalion commanders to gather information about what types of personal problems they have found as leaders and what solution techniques were employed to deal with the problems. Leaders' attitudes toward soldiers were also obtained as a basis for differentiating between developmental, punitive leaders, and "administrative" leaders. Subjects were 326 male battalion commanders assigned to locations throughout the continental US, Europe, and the Far East. Responses regarding soldier problems were submitted to a principle component, varimax rotation factor analysis. The four generic factors identified were (1) on the job problems, (2) substance abuse problems (3) socio-emotional problems, and (4) AWOL. The three leadership styles differed most in their handling of job performance problems, less regarding socio-emotional problems, and less yet on substance abuse problems. There were no discernable differences regarding AWOL.

Relationships Between Attitudes and Leadership Style:
A Policy Capturing Approach

Jerome Adams
Associate Professor
Science Research Lab, USMA
West Point, NY 10996

John Richards
Assistant Professor
D/BS&L USMA
West Point, NY 10996

Terrence Fullerton
Research Psychologist
Cadet Counseling Center, USMA
West Point, NY 10996

Objectives of the Study

The main objective of this study is to attain more understanding of the relational qualities of leadership in military organizations. These qualities are often described in terms of leader style as in Fiedler's contingency model (1964); however, the emphasis is on a more focused view of "style" as it relates specific leader behaviors to standard organizational problems. The previous failure of most traditional personality measures to predict leader behavior has caused many scientists to adopt the newer view that such behavior is an interactive function of personality, dispositions, and the situation (Hollander & Neider). What seems necessary to further the predictive quality of the study of leadership in groups and organizations is a better grasp of these interactive qualities in a standardized social interaction from a policy-capturing perspective. Therefore, the present study seeks to extend the understanding of the relational qualities of leadership using policy capturing. This approach requires analyzing the styles of leadership used by commanders to deal with problem soldiers. Several important factors give special meaning to this study:

* The research report here was supported by Grant #13 ARI 80-32 from the U.S. Army Research Institute for the Social and Behavioral Sciences.

** This paper represents the views of the authors and not the official position of the U.S. Military Academy, the U.S. Army Research Institute, the U.S. Army, or any other governmental agency unless so designated by authorized documents.

*** We thank the staff of the ARI field unit, Fort Benjamin Harrison, Indiana for their assistance in collecting and coding the data.

- Commanders are senior leaders in organizations whose decisions directly influence outcome measures of subordinate members.

- The study uses four standard categories of leader problems which all military commanders must face.

- The study uses fourteen standard solution techniques to resolve the four problem categories.

- The study uses a standard personality attitudinal measure.

- The study was conducted in the real-life context of more than three hundred organizations with leaders performing similar organizational functions.

The formal and objective approach of policy capturing in leadership provides a conceptual framework of how leaders behave in a given situation. For example, we will determine if each military leader is considering the same information and placing the same importance on specific problem solutions for a given set of standard problems. Only a small portion of research concerning policy capturing has been concerned with industrial problems (see Christal 1968; Naylor & Wherry, 1965; Christal 1969).

Although regrettable, the paucity of research is due to practical constraints. Policy capturing methods require a large number of comparable judges; hence, the methods are typically applicable only in large homogeneous organizations. However, a setting such as the military would be an ideal setting to determine how similar leaders are in their behaviors toward standard leadership problems.

Method

As indicated, the research to be reported is an attempt to study leadership as an interactive process using a policy capturing perspective. The technique used required the administration of a questionnaire to commanders to gather information about what types of problems they have found as leaders and what solution techniques were employed to deal with the leadership problems. Information was also obtained about the leaders' attitudes toward problem soldiers as a basis for differentiating between developmental leaders, punitive leaders, and mere administrative adjudication processing.

Subjects

Subjects were 326 Army commanders assigned to locations throughout the continental United States. The research was part of a larger program geared to developing more effective unit-level techniques for addressing the problems among first term enlisted soldiers.

Procedure

Military leaders were tasked to complete a questionnaire about their attitudes toward problem soldiers and what leadership style they used to resolve problem issues. The information contained in this study is restricted to the military leaders' responses to questions of working with problem soldiers.

In an effort to reduce the number of categories of soldier problems rated in the survey, it was decided to submit the leaders' responses to questions about problem soldiers to a principal component, varimax rotation, factor analysis. This analysis suggested that four factors be retained. The four retained factors were thus conceptualized as: I - On the Job Problems, II - Substance Abuse Problems, III - Socio-Emotional Problems, IV - AWOL. Together the factors accounted for 55.7% of the observed variance. These dimensions permitted logical groupings of the original 14 problems into four supra-ordinate, broader problem areas. These clusterings were then imposed on two related survey sections dealing with the perceived frequency of soldier problems and the estimated likelihood of problems ultimately resulting in separation from service.

In order to decrease the number of solution categories to problems, it was determined to factor analyze the solutions to four problems: job performance, substance abuse, marital problems and AWOL. These four problems were chosen because they best represented (by virtue of factor loadings) the four supra-ordinate categories of soldier problems. This factor analysis and rational consideration suggested the following further supra-ordinate solution categories: 1) Informal Counseling; 2) Non-punitive Aids to Military Adjustment; 3) Remedies for Substance Abuse Problems; 4) Punitive Non-discharge Remedies; 5) Early Discharge Program; 6) Separation for University and 7) Other Discharges. The next set of analyses describe the attempt to determine predictors of personnel management style. Using commander's demographic attitudinal and personal characteristics as predictors these variables were entered into a series of multiple regressions to see how much they predicted leader style. Leader style is a criterion measure for each commander. This score is computed by creating a 7 x 4 matrix. Thus, each commander has a score in 28 cells of this solution/problem matrix. The proportion of variance accounted for by the predictor variables was as expected, minimal $R^2 = .013$ to $R^2 = .109$.

The next set of analyses allowed us to examine the relational qualities of traits, behavior, and situation from a policy capturing perspective. In order to accomplish this, a series of discriminant analyses were performed. The results are given in Table 1.

The results of all of the analyses are described in more detail in the full paper. Because of space limitations, we will only summarize our findings as the data edify the complex relational qualities of leader attributes, style and work situation. First, the attempt to predict leader behaviors from personality information was not successful. This outcome is consistent with early research by the trait theorists who tried to characterize effective versus ineffective leaders based upon personal characteristics. Among the characteristics of the leaders studied here, the best predictor of leadership behavior is attitude toward problem soldiers. Clearly, it is that characteristic which is most clearly important and salient with the situation. Sherif (1948) noted that results such as those presented here, when he observed that leadership is not determined by "personal qualities in the abstract" (p. 456). Finally, these results are also consistent with Stogdill's (1974) observation; leadership performance is most often determined by a pattern of personal characteristics in a situationally specific setting. Hence, contributing personal characteristics are dependent upon the situation.

Such preliminary findings have led us in this research to consider the relational qualities of personal characteristics with a behavioral outcome using homogeneous problems in similar role level settings as exemplified in the highly structured organization found in the military.

Consistently, we were able to significantly discriminate between leaders whose behavior toward standard problems was developmental or punitive and in the last case merely administrative. Generally speaking, we found support for differences in how military leaders behave to the same problems. The leaders who were classified as developmental were most distinct in how they behaved as evidenced by the soltuions they chose to deal with problem soldiers. This suggests a high degree of similarity of value or judgement among developmental leaders in terms of consistency of behavior when addressing realistic, standard problems in common organizational settings.

The perspective of policy capturing serves as a useful method in revealing that overall, developmental leaders make similar judgements on appropriate behavior for areas where the problem soldier can be given the opportunity to improve and possibly make long-term meaningful contributions to the organization. These leaders' actions are highly consistent to underscore a value system that leads to judgemental consistency of their behavior. Obviously, punitive leaders, who do not espouse the same personal developmental attitudes as a dispositional characteristic, differ from those who do in terms of the degree to which the relational qualities of traits and styles affect decisions for standard problems in homogeneous settings. Where we expected to find the least difference between developmental and punitive leaders (in the administrative choice outcomes) the evidence substantiates this similarity.

Although possible alternative explanations can never be ruled out, it is possible that the train-style-situation leadership link is actually due to some extraneous factor. However, the work of Kelley (1972) suggests three major dimensions for such an intuitive examination. If the evidence does not hold up over time or across relevant situations, or if it is not supported by the opinions of other relevant actors, leader phenomenon may plausibly be attributed to some exogenous source. In our study, the results were based upon leader style and situations for periods ranging up to eighteen months, across more than three hundred similar levels of authority governed by the same standard procedures of the organization's hierarchy. It is not known at this time how the leaders, themselves, would support the conclusions of this study. However, if we accept that these findings accurately reflect the relational qualities of leader traits and styles to a standard set of problems in the same situation, these results suggest that policy capturing is a conceptually meaningful methodology to better explain the process of leadership in specific organizational settings.

1. Christal, R.E., "JAN: A Technique for Analyzing Group Judgement" Journal of Experimental Education. 1968, 36, 24-7

2. Fiedler, F.E., "A Theory of Leadership Effectiveness." New York: McGraw Hill, 1967.

3. Fivars, G., "The Critical Incident Technique". A Bibliography. Palo Alto, CA; American Institutes for Research, 1973.

4. Fleishman, E.A., "The Leadership Opinion Questionnaire" in R.M. Stogdill & A.E. Coons (eds) Leader Behavior and Its Description and Measurement. Columbus: Ohio State U., 1957

5. Hemphill, J.K., & Coons, A.E., "Development of the Leader Behavior Description Questionnaire" in R.M. Stogdill & A.E. Coons (eds) Leader Behavior Its Description and Measurement. Columbus: Ohio State U., 1957.

6. Hollander, E.P., & Julian, J.W., "Contemporary Trends in the Analysis of Leadership Processes" Psychological Bulletin. 1969, 387-97

7. Hollander, E.P., & Neider, L.L. "An Exploratory Study Using Critical Incidents and Rating Scales to Compare Good & Bad Leadership," Buffalo, NY: S.U.N.Y., Tech Rept. #5 July 1977.

8. House, R.J., "A Path-Goal Theory of Leader Effectiveness" Administrative Sciences Quarterly 1971, 321-32.

9. Hunt, J.G., & Tarson, L.L., Leadership: The Cutting Edge. Carbondale, Ill: Southern Illinois U., 1977.

10. Kelley, H., Causal Schemata and the Attribution Process. Morristown, N.J.: General Learning Press, 1972.

11. Madden, J.M., "An Application of the Policy Capturing Method" Bulletin of the Psychomic Society. 1975, 6, 619-21.

12. Naylor, J.C., & Wherry, R.J., Jr., "The Use of Simulated Stimuli and the JAN Technique to Capture and Cluster the Policies of Raters, "Educational and Psychological Measurement. 1965, 25, 691-92

13. Nie, N.H., Hull, C.H., Jenkins, J.G., Steinbrunner, K., & Bent, D.H.; SPSS New York: McGraw Hill Publishers, 1975.

14. Sherif, M., "An Outline of Social Psychology" New York: Harper & Rolo, 1948.

15. Stogdill, R.M., "Handbook of Leadership". New York: Free Press, 1974

16. Stogdill, R.M., & Shartle, C.L., "Methods for Determining Patterns of Leadership Behavior in Relation to Organization Structure and Objectives" Journal of Applied Psychology; 1948, 32, 286-91.

17. Wheeler, H.N., "Punishment Theory and Industrial Discipline" Industrial Relations; May 1976, 15, 235-43.

18. Zedeck, S. & Blood, M.R., "Foundations of Behavioral Science Research in Organizations." Belmont, C-A: Brooks/Cole, 1974.

TABLE I

CANONICAL DISCRIMINANT FUNCTIONAL
On Leader Style For Composite Job Problems, AWOL
Socio-Emotional and Substance ADVSE

Function	Eigenvalue	% Variance	Canonical Correlation	Wilks' Lambda	CH 1 SQ	D.F.	Significance
1	.04	100	.20	.959	13.23	4	.01

STANDARDIZED DISCRIMINANT FUNCTION COEFFICIENTS

Job Problem	-.57
Substance ADVSE	-.13
Socio-Emotional	-.50
AWOL	.30

Group Centroids

Group	
1	-.21
2	.19

Affourtit, Thomas D., Interaction Research Institute, Inc., Fairfax, Virginia. (Wed. A.M.)

The Leadership Evaluation and Analysis Program (LEAP): Validity and Future Directions

The LEAP is a self-applied OD method that allows small unit commanders to assess leadership concerns, measure unit combat readiness, and evaluate decision-making effectiveness. The self-development strategy assures individual command control and confidentiality, while central analysis is made possible through voluntary and anonymous submission of data.

Production rates, reenlistments, and absenteeism were used to validate LEAP Interaction Inventory scales and a Disparity Index (DI) was developed as an additional Dimension of command climate to measure disunity and differential treatment of unit members. The DI, the strongest predictor of performance, is the subject of planned research in working relationships using the Vertical Linkage Dyad model of analysis.

THE LEADERSHIP EVALUATION AND
ANALYSIS PROGRAM (LEAP)

Validity and Future Application

Thomas D. Affourtit
Interaction Research Institute, Inc.

The Leadership Evaluation and Analysis Program, commonly known as the LEAP, was originally designed for the U.S. Marine Corps.

- The LEAP is designed for use by the small unit commander to identify leadership concerns, to measure overall unit combat readiness, and to evaluate the effectiveness of the decision-making process.

- The LEAP provides company, battery, and squadron level commanders with a completely decentralized leadership aid. The program is self-applied and voluntary on the leader's part. And, the results are strictly confidential to the individual command.

- Most important, through the decision-making feedback principle, the LEAP aids the leader in developing the flexibility to deal with various groups in a variety of situations and mission requirements.

The LEAP is basically an intelligence-gathering process, founded on the principles of modern management methods, and utilizing behavioral science techniques. The entire program is presented in a manual that features a step-by-step procedure for administering the techniques and explicit guidelines for scoring and reviewing the results. Therefore, no outside professional assistance is necessary and no reports or formal paperwork is required.

Program materials allow the commander to systematically measure unit performance in terms of general Marine Corps standards and specific unit requirements. In addition, unit motivation is measured to determine the reasons behind various levels of performance. Since performance is the consequence of a motivational state, the LEAP attempts to measure the causes as well as the effects of unit combat readiness.

Once the causes of performance levels are identified, the leader can take the necessary action to extinguish those conditions that produce deficient behavior, and reinforce or support the conditions that promote positive performance.

The technique for measuring unit motivation is a questionnaire that functions much like a starlite scope. It enables the leader to see the hidden causes of performance, which are not easily observed under normal conditions. In this way, the LEAP procedure is not unlike counterinsurgent or search-and-destroy operations. But, in this case, the enemy is indifference, negligence, discord, and prejudice. These are some of the reasons that separate effective, combat-ready units from ineffective ones.

The application of the LEAP is very simple, requiring only a few minutes of the commander's time to request application, to review the results, and then to make appropriate counteractive decisions.

First, the commander has the unit clerk record various performance statistics on the Leadership Analysis Form, a behavioral measuring technique and an essential part of the program. This part is easily accomplished since all the information needed is available from unit records. It takes about an hour to accomplish.

Next, the CO directs a responsible subordinate leader to administer the Interaction Inventory, the motivational questionnaire, to the entire command to measure the motivational level of several relevant areas of Marine Corps concern. It takes from 15 to 35 minutes for troops to complete the anonymous questionnaire. Questionnaire results may be either scored manually or computer processed.

Finally, the CO reviews the results of the questionnaire according to a scheme outlined in the manual, or according to his or her own interest. These results serve as benchmarks for the command which are used to judge progress in critical leadership areas.

Now, let's take a closer look at how the LEAP works.

The CO calculates his or her own performance profile, a procedure that is similar to developing the readiness indicators used at some higher level commands, except that unit leaders select their own areas of importance, and unit measures are taken over a designated period of time.

The command motivational profile is then reviewed to determine the areas of strength and weakness in terms of troop perceptions. Several general areas are measured, such as command efficiency and cohesion, which indicate levels of unit preparedness. And, conditions of discrimination, justice, and intergroup climate provide an assessment of overall command equality.

In addition, two adjunct survey instruments have been developed to expand the domain of the motivational data and to provide flexibility of application through selection of scales considered most relevant at the user unit level. Adjunct scales include measures of senior proficiency, senior support, communication flow, organization and planning, recognition, discipline, task satisfaction, task significance, functional readiness, solidarity, and individual development.

Like an aerial reconnaissance photo, this information can be amplified to expose specific issues and conditions that make up each scale in the command motivational profile. Again, strengths and weaknesses are noted as peaks or depressions along a scale graph. Tactically speaking, this information gives the coordinates of the enemy's position.

These data also allow the leader to establish mission objectives or goals to reach in order to destroy the negative condition and improve unit military status.

Validity studies conducted on Marine Corps field units show that as these scale scores move toward the right, indicating more positive motivational levels, unauthorized absences within a command significantly decrease, and first-term reenlistment rates significantly increase.

The motivational information can also be analyzed from the standpoint of any number of groups within the command, such as senior vs. subordinate Marines, minority vs. majority Marines, or by educational level, marital status, and career intention groups, just to name a few.

It has been found that where disparity exists between groups distinguished by rank, such as sergeants and below vs. staff NCO's and above, absenteeism will again be high and reenlistments will be low. But, when the scores of the two groups move closer together, indicating that both groups judge the command as functioning at the same level, unauthorized absences again decrease significantly and reenlistments increase significantly. The same outcome was discovered for differences in perception between minority and majority member troops.

In review, there are just four primary steps in the LEAP procedure:

- First, identify leadership problem areas or locate enemy positions using the LEAP tactical sensing and reconnaissance devices (analyze).
- Second, establish leadership goals and management objectives (plan).
- Third, take corrective action (attack).
- Fourth, evaluate results in terms of performance and/or motivational outcomes (evaluate).

In this same manner, any leadership decision, program of training, or policy designed to improve performance can be evaluated. Such decision-making evaluation assists Marines in producing viable solutions to some of the contemporary problems that all Marines face today.

In support of this solution-oriented approach, an information-processing data bank has been initiated. The LEAP Network Monitor System is a process that will enable Marines to share solutions to common problems encountered in the field. This system will function on a voluntary and an anonymous basis as a data input and inquiry bank for Marines using the LEAP. Information input, recommendations, and solutions discovered can be analyzed, and data feedback can be presented in consideration of any number of influencing conditions, such as unit composition, mission, unit status, location, or effective strength. Such results, based on actual conditions, will also be beneficial in training new leaders to make the most appropriate and effective decisions prior to command assignment.

In essence, the LEAP supports one of the most basic and time-proven leadership principles -- know your personnel! It doesn't matter whether your leadership style is authoritarian, participatory, or charismatic; you can benefit from a better understanding of unit personnel and more systematic knowledge of how members react in various situations.

In summary, the LEAP is a voluntary aid for company level commanders. And the payoff for just a little time and effort invested can be:

- Greater unit harmony and morale.
- Performance improvement.
- Less disciplinary effort expended.
- Increased control and influence over troops.
- Leadership skill development.

Validity studies conducted over a five year period revealed significant relationships between various LEAP scale scores and unit reenlistment and unauthorized absence rates, first term and career reenlistment intention, drug and alcohol abuse, and theft within participating commands. In addition, a longitudinal study of avaition maintenance commands showed that LEAP motivational/ climate measures correlated significantly with such productivity indices as defect rates, parts and maintenance backlog, and inventory readiness.

One of the major outcomes of the LEAP studies was the development of the disparity index (DI) as a measure of differential working relationships between pertinent groups within commands. Beyond measuring the degree of unity within a command over critical conditions, the DI was a more efficient predictor of climate and performance for some of the LEAP scales.

A proposed future effort involves using the LEAP methodology to investigate the impact of internal role relationships upon unit efficiency. Intraunit differentiation as measured by the DI is a common condition that has a significant impact upon group performance. While traditional attempts to study leadership focus on average style and treats within group variance as error variance, the DI model considers intragroup variance as valid. Any serious consideration of unit cohesion, adaptation, and internalization of military norms must take into account that leaders differ over time and within groups and that groups respond differently.

As a flexible and valid organizational development intervention and research tool, the LEAP provides field commanders with a procedure for improving unit combat readiness, and offers scientists a unique method of data gathering and analysis to study the complex phenomena called leadership.

REFERENCES

Affourtit, T.D. Leadership self-development: A new approach toward organizational efficiency (Professional Paper 76-8). Fairfax, Va.: Interaction Research Institute, Inc., November 1976.

Affourtit, T.D. The Marine Corps' Leadership Evaluation and Analysis Program (LEAP) volume I: Manual for the Leadership Analysis Form and LEAP Interaction Inventory (2nd ed.) (Technical Manual 77-1). Fairfax, Va.: Interaction Research Institute, Inc., March 1977.

Affourtit, T.D. The Leadership Evaluation and Analysis Program (LEAP) volume II: Development and validation of the Interaction Inventory (Technical Report 78-3). Fairfax, Va.: Interaction Research Institute, Inc., April 1978.

Affourtit, T.D. Dimensions of organizational and management development: An operational model (Professional Paper 79-1). Fairfax, Va.: Interaction Research Institute, Inc., January 1979.

Affourtit, T.D. The Leadership Evaluation and Analysis Program (LEAP) Volume II: Validation update (Technical Report 79-3). Fairfax, Va.: Interaction Research Institute, Inc., March 1979.

Affourtit, T.D. The Leadership Evaluation and Analysis Program (LEAP) manual for the Interaction Inventory adjunct questionnaires (Technical Manual 79-6). Fairfax, Va.: Interaction Research Institute, Inc., May 1979.

Affourtit, T.D. The Leadership Evaluation and Analysis Program (LEAP). Economic feasibility report (Technical Report 79-8). Fairfax, Va.: Interaction Research Institute, Inc., July 1979.

Affourtit, T.D. The Leadership Evaluation and Analysis Program (LEAP). Contemporary Marine Corps leadership issues: Final report 1979 (Technical Report 79-10) Fairfax, Va.: Interaction Research Institute, Inc., June 1979.

THE JOB TASK ANALYSIS/SKILLS AND KNOWLEDGE MARRIAGE (PART 1)

T. M. Ansbro and W. A. Hayes
CNET/TAEG NEPDIS Development Team, CNET HQ. NAS Pensacola

As job/task analysis methodology continues to advance in sophistication, the computer takes over an ever greater share of the work of analysis, leaving less and less of formerly judgemental areas to assumption. But, assumption still functions where job "skills and knowledges" are assigned as underlying or component to tasks in inventories.

So far, in front-end analysis of the workplace, task interrelationships, rankings by complexity, and degrees of commonality can be readily determined by the computer. If task 00165 in package 017 proves common to 16 others in an inventory of 2500, has a complexity index of 1.25, and embodies all the subordinate work behaviors of 137 other tasks, the computer can record these features and position the subject task appropriately in an output hierarchy. It can also sort on the basis of identifying or descriptive data included in the task record in the inventory. Such processing gets pretty far down into specifics of work behaviors underlying tasks, but it doesn't affix identified manipulative or processing skills and static-descriptive or process-associated knowledge (information) elements to those tasks.

This paper (Part I) describes a matrix of "skills and knowledge" elements to augment a model front-end job/task analysis subsystem (NEPDIS--Naval Enlisted Professional Development Information System) and discusses such alternatives as adding these data to the master job/task inventory or providing an ancillary "skills and knowledge" inventory for use of the training program developer.

As "front-end" job/task analysis methodologies have progressed and continued to advance in sophistication, the computer takes over an even greater share of the work of analysis, leaving less and less of formerly judgemental areas to panel-of-experts analysis, evaluation, and cataloguing. Identification and description of tasks in inventories appear fairly concrete, as they do for such task component work behaviors as task elements (Johnson and Richmann, 1975). What analysis is possible so far with the task (and its accompanying descriptive factors) as the sole data source has yielded task complexity indexes, hierarchies expressive of task interrelationships, and task commonality indicators within and among occupational fields. With such outputs of job/task analysis producible by computer programming, assumption and subject-matter expert (SME) consensus might well be expected to fade into the background. However, assumption still functions where such work behaviors as job "skills and knowledges" are or must be assigned as underlying or component to tasks in inventories.

In Navy manpower management, ship and squadron manning documents and job (billet) descriptions are dependent in the main upon extensive, detailed, and comprehensive inventories of operational, maintenance, administrative, military watchstanding, and other tasks (as well as ship, systems, and equipment data). Personnel distribution, rating assignment, advancement, and training (especially training) depend upon inventories of skills as well as tasks; and the training community needs to take the job/task inventory "audit trail" down farther still--to the level where "job knowledge" can be associated directly with "job skills" to support job tasks.

This paper (Part I) describes an attempt to produce a matrix of "job skills and knowledges" elements to augment model front-end job/task analysis subsystem currently under development by the staff of the Chief of Naval Education and Training (CNET) and Training Analysis and Evaluation Group (TEAG) elements in Pensacola, Florida. The subsystem model is the Training Analysis Subsystem of the Naval Enlisted Professional Development Information System (NEPDIS) (Davis, 1976, 1977, 1977a, 1977b).

The NEPDIS model currently stores some 23,000 Naval Avionics rating tasks in its inventory. Occupational data acquisition was accomplished for other ratings in the Naval aviation community, ten in all, but these data are not yet in the computer. As a result, occupational data stored and analyzed to date by NEPDIS remain in the major functional category of Maintenance. A typical listing of an avionics maintenance task is shown in Figure 1.

With computerized analysis of such data as typified by this entry in the Aviation Electrician's Mate (AE) job/task inventory under NEPDIS, tasks may be ranked by complexity; and degrees of commonality (from the identical to an agreed-upon level of similarity) (Davis and Perry, 1980) may be established. Task interrelationships may also be established. Some tasks obviously contain many component work behaviors that are also contained in other tasks of lesser complexity and scope; some tasks duplicate the work behaviors of others, sometimes regardless of the task titles involved. Tasks shown to "embody" other (subordinate or component) tasks are termed "Omnibus" tasks; the tasks shown to be component to the Omnibus tasks are termed "Embodied" tasks (Figure 2). In the NEPDIS job/task inventory (JTI) AE task 00165 in package 017 may be shown to have a complexity index of 1.25, prove common to 28 others in a one-rating inventory of 3700 tasks, and embody all the component work behaviors of 137 other tasks. The computer can record these features after producing them via analysis, and it can position the subject task appropriately in any specified output hierarchy. It can also sort on the basis of identifying or descriptive data included in the task record in the inventory (Ansbro, 1978). Such processing gets pretty far down into the specifics of work behaviors underlying tasks, but it doesn't extend beyond identifying job-related skills. Figure 3 shows task "signature block" (work-behavior descriptive data) printed out. The five skill areas included in the task signature block contain statements of work behavior that would appear to be as descriptive of elements (or components) of a task as of the skills that they represent. They are definitive, small in compass, and specific to (and therefore underly) task performance; therefore attached to other descriptive data for the task.

Herein lies a problem for the training program developer or curriculum designer. To design a training course, he needs an inventory of tasks to describe course graduate job performance capabilities and to provide realistic practical exercises and performance tests. Successful student/graduate task performance when matched with on-job (billet) requirements in the fleet (also tasks) serves as a reasonable predictor of successful performance on the job.

However, can a course cover all the tasks that the graduate must perform on the job in his fleet assignment? The best that we can hope for is coverage of these tasks that best represent fleet requirements. The analysis that results from this realization requires intricate grouping and cataloguing of work behaviors. Selection of representative tasks really involves selection of those underlying behaviors component to or most widely transferable among tasks assumed to be representative of fleet job requirements.

The transferable component/supporting work behaviors underlying task performance are skills. Skills, stated in behavioral action language, resemble tasks. Indeed, the workplace and the schoolhouse both use task and skill terminology almost interchangeably. As an example, welding is described as a transferable skill, since welding something to something else is component to performing many metal fabrication and repair tasks. However, depending upon how a work-behavior statement reads ("weld fire-hose support bracket to bulkhead"), welding may be regarded as the action part of a task statement. Welding as a major work behavior can also be regarded as more encompassing than a task; it can represent an entire worker career, or the sole mission or output of a shop or department. Soldering, somewhat similar technically, but a considerably smaller skill, is usually termed just that—a skill. Viewed from a task-descriptive orientation, it is also a task element. But, because of its simplicity, subordinate/component nature, and wide applicability (therefore transferability) to task performance, it is generally considered to be a skill, and in the occupational field(s) of electricity/electronics, a basic one, at that.

On the premise that if the schoolhouse is to train the graduate to perform on the job in the fleet, the instructional program/course designer must attempt to replicate the best representative and most critical tasks from that target environment; then, he must ferret out, verify and train on those skills found to be component to those tasks selected for training and transferable to those known to exist in the workplace but not selected for training. Therefore, the training community needs extensive and comprehensive JTIs with task-descriptive data as fully fleshed out as possible; and it is certainly more than merely convenient to have skills appropriately identified as such. Further, in any career continuum, it must follow that in the earlier training programs (basic, apprentice, initial job-entry), the concentration of training effort is on skills, transferable to the job environment where they may be applied and refined in a work-and-OJT setting. Advanced specialized training still teaches skills, but task performance figures somewhat more prominently, as higher level technician training more closely approximates the real-world environment of the graduate's ultimate work site.

As mentioned earlier, skill statements resemble task (and task element) statements: There are action verbs, objects of action, and job-environment conditions and work-performance standards. It is necessary to make one clear distinction if there is to be any observable difference between these statements (in an inventory). In the task statement, the object of the action is specific: a clearly identified or coded item (system or subsystem component, equipment item, part, form, machine, instrument, etc.) In the skill statement, the object of the action is not a specific item; it can be typical of a group or class, a "generic" item (mild steel plate, galvanized sheet metal ducting, bar stock, tubing, circuit wiring, solid state printed circuits, etc.), even a synthesized or composite item generated for such a purpose as training or practice of an identified skill action. As tasks fall into hierarchies, so also do skills. A "troubleshooting" task (in NEPDIS: "Isolate Fault/Troubleshoot --------object") employs subordinate (component) "troubleshooting" skills: selecting/using references, selecting/using tools, selecting/using support materials, selecting/using support equipment, and selecting/using test equipment.

The principal mechanistic reason for making these task/skill distinctions in the NEPDIS Training Analysis subsystem is the need for the computer to recognize the distinctions in its progress through analysis toward such outputs as billet-specific task inventories and rating-specific skill inventories. NEPDIS front-end analysis was designed to be totally computer-served and to conduct all job/task/skill analysis for designated users at the "front-end"; hence, the emphasis on coding, detailed identification and descriptive factors, and other aspects of an audit trail from task identification through reference source.

Job knowledge, or the task/skill performance enabling base (information/data); lies at the bottom of the audit trail. Information from all reference sources pertinent to task/skill performance can be assumed to fit into a relatively simple matrix, an example of which is shown in Figure 4. A NEPDIS-conducted literature search based on reference text support of occupational data already in the JTI indicated substantial reference support for the details of task and task element performance. However, little test support for those behaviors identified as component skills was found in these references. For instance, what to solder and at what point to solder, and what tools and materials applied to the task element were amply covered by reference. How to solder was not. Hence, in a Navy-designed front-end job/task analysis subsystem built to support Instructional Systems Development (ISD), the subsystem developers discovered that in some instances they had provided themselves with relatively light direct reference text support for developing the essentials of some skills training. By tracking back through the appropriate job task-oriented references and by recourse to rating-specific texts and existing skills-training school texts, the necessary reference support can be found. But it is not direct, and it is not totally and specifically contained within the master JTI for the rating or occupational field.

> NEPDIS front-end analysis had been designed with the avowed intent of keeping all the job/task (and skill) analysis at the front end. An instructional systems developer was to receive all the various inventories (subsystem outputs) needed to develop curricula/instructional programs, etc. without having to "go back to the front end" himself to analyze or further analyze data; especially, he should not have to collect raw data.

One simple solution is to add such text references in the appropriate spot in the task identification block in the master JTI. Another is to provide a brief structured addendum to the master JTI, this item strictly for the use of training program development personnel. Figure 5 illustrates the general scope of basic supporting information useful to the developer of skills training. In essence, this example would suggest the beginning of an adjunct task/skills performance-supporting information inventory or skills and knowledge library.

A third alternative is to construct a matrix such as the one shown (in concept) in Figure 4, and code it to the task signature block in the JTI. The matrix generally illustrates the task support hierarchy: from the top down, task performance is supported by any number of task elements; the task elements are supported by (and incorporate) manipulative and/or information-processing skills; and these skills are supported by static descriptive and process-associated knowledge items (the anabling information base). For practical incorporation into the JTI, identified skills can be cross-coded to task identification codes, and bodies of information identified as skill-supporting items.

The alternatives mentioned above represent current NEPDIS effort to marry the already definitive job task information in the JTIs with equally definitive supporting skill and knowledge data. The intent also is to maintain a visible and easily followed audit trail throughout the Training Analysis Subsystem.

REFERENCES

Ansbro, T. M. "Using the Computer to Build the Task Inventory." Proceedings - 20th Annual Conference of the Military Testing Asociation, Oklahoma City, OK, 1978, 263-274.

Davis, D. D. "Dialogue." CAMPUS, The Naval Education and Training Monthly. January, 1976, 24-27.

Davis, D. D. "Dialogue." CAMPUS, The Naval Education and Training Monthly. April, 1977, 22-27.

Davis, D. D. "Naval Enlisted Professional Development Information Support System." Proceedings of the National Security Industrial Association on the State-of-the-Art Application of Advanced Training Technology. April, 1977a.

Davis, D. D. "Naval Enlisted Professional Development Information Support System." Journal of Educational Technology Systems, 1977b, 6(2), 133-141.

Davis, D. D., Perry, N. N. "Determining Task Commonality in Navy Training-Two Methods Examined." Proceedings - 22nd Annual Conference of the Military Testing Association, Toronto, Ontario, Canada, 1980, 229-237.

Johnson, R. N., Richmann, J. N. "Task Analysis: The Basis for Criterion Test and Curriculum Design." Proceedings - 17th Annual Conference of the Military Testing Association, Indianapolis, IN, 1975, 538-550.

RATING = AE   PACKAGE = 0001   TASK = 0020   DUTY SUBCATEGORY = 01

TASK STATEMENT = PERFORM MAINTENANCE CHECK ON ROTATING BEACON ASSEMBLY 44126

PLATFORM = P-3 A/B (APRC)
SYSTEM   = LIGHTING SYSTEM (44000)
EQUIPMENT = EXTERIOR LIGHTING (44100)
COMPONENT = ROTATING BEACON ASSEMBLY (44126)

COMPLEXITY = 1.67

MAJOR ACTION CATEGORY = MAINTAIN
DUTY SUBCATEGORY (01) = CHECK/TEST/INSPECT
TASK ACTION (PMC)   = PERFORM MAINTENANCE CHECK

CUE..........................REPAIR COMPLETE
STANDARD.....................IAW REFERENCE PUBLICATION
REFERENCE....................NA-01-75PAA-2-12
TOOL.........................COMMON HAND TOOLS, SPECIAL HAND TOOLS
SUPPORT MATERIAL.............FUSE/SWITCH
SUPPORT EQUIPMENT............POWER UNIT 28V 400HZ
TEST EQUIPMENT...............RPM GAUGE, AMMETER, MULTIMETER

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
GENERAL..........................	2	1	1	3	1	1	1	1	2	0	0	0	0	0
DUTY SUB 01......................	2	3	2	0	1	3	1	0	0	1	0	0	0	0
SKILL 1 (REFERENCE)..............	2	1	1	1	1	0	0	0	0	0	0	0	0	0
SKILL 2 (TOOL)...................	3	3	2	1	0	0	0	0	0	0	0	0	0	0
SKILL 3 (SUPPORT MATRL)..........	2	0	0	0	0	0	0	0	0	0	0	0	0	0
SKILL 4 (SUPPORT EQUIP)..........	1	1	1	1	0	3	0	0	0	0	0	0	0	0
SKILL 5 (TEST EQUIP).............	3	2	2	3	2	2	2	1	3	0	0	0	0	0

FIGURE 1.   NEPDIS TASK LISTING: AVIONICS; TASK IN AVIATION ELECTRICIAN'S MATE JOB/TASK INVENTORY

```
AE 0022 0002 01 IFT M4 ISOLATE FAULT/TROUBLESHOOT AFCS 57560
GENERAL     2131112210000000000000000000   COMPLEXITY = 2.39                137
D.S.     1  1331331000000000000000000000   AE 0022 0002 01 IFT M4
SKILL    1  2111300000000000000000000000
SKILL    2  1111000000000000000000000000
SKILL    3  3000000000000000000000000000
SKILL    4  3333331000000000000000000000
SKILL    5  3333332220000000000000000000

AE 0015 0222 01 POC C3 OPERATIONAL CHECK SW RATE GYRO CN 495A/AJB-3 57A4020
GENERAL     2213211271000000000000000000   COMPLEXITY = 2.38                138
D.S.     1  3323332323000000000000000000   AE 0015 0222 01 POC C3
SKILL    1  2221300000000000000000000000
SKILL    2  2322000000000000000000000000
SKILL    3  3233223230000000000000000000

AE 0015 0168 01 POC M4 OPERATIONAL CHECK DISPLACEMENT GYRO 57A41
GENERAL     2213211220000000000000000000   COMPLEXITY = 2.36   EMBODIED 111   1   139
D.S.     1  3323332323000000000000000000   AE 0015 0168 01 POC M4
SKILL    1  2221300000000000000000000000
SKILL    2  2322000000000000000000000000
SKILL    3  3233223230000000000000000000

AE 0008 0098 01 PMC C3 PERFORM BENCH CHECK AUTO PILOT WARNING LIGHTS FLASHER 57110
GENERAL     2113211120000000000000000000   COMPLEXITY = 1.34   EMBODIED 112   2   140
D.S.     1  2323210003000000000000000000   AE 000 0098 01 PMC C3
SKILL    1  2111300000000000000000000000
SKILL    3  1213211130000000000000000000

AE 0024 0646 01 IFT C3 ISOLATE FAULT/TROUBLESHOOT MT 3409 57578
GENERAL     2111111210000000000000000000   COMPLEXITY = 1.02   EMBODIED 113   3   141
D.S.     1  3313111002000000000000000000   AE 0024 0646 01 IFT C3
SKILL    2  2311000000000000000000000000
SKILL    3  1223211210000000000000000000

AE 0022 0002 01 IFT M4 ISOLATE FAULT/TROUBLESHOOT AFCS 57510
GENERAL     3123100020000000000000000000   COMPLEXITY = 2.37
D.S.     1  3333333003000000000000000000   AE 0022 0002 01 IFT M4           142
```

OMNIBUS

FIGURE 2. OMNIBUS/EMBODIED TASK RELATIONSHIP (AVIATION ELECTRICIAN'S MATE INVENTORY)

SKILL AREA: (1) USE REFERENCES

3. LOCATE DATA (WITHIN REFERENCES)

   A. INFORMATION IS INDEXED AND SEQUENTIALLY ARRANGED WITHIN ONE REFERENCE.

SKILL AREA: (2) USE TOOLS

1. SELECT TOOLS (TYPE AND SIZE)

   C. ONE SPECIFIC TYPE AND SIZE MUST BE USED. A SUBSTITUTE CANNOT BE USED; E.G., TOOL FOR ADJUSTING THE IRON CORE OF AN INDUCTOR, SPANNER WRENCH.

SKILL AREA: (3) USE SUPPORT MATERIALS

1. SELECT SUPPORT MATERIALS (SPECIFICATIONS)

   B. SPECIFICATIONS GIVEN; SEVERAL SUBSTITUTES PERMITTED; E.G., CLEANING SOLVENT (FED. SPEC. NO. P-S-661 OR PD-680).

SKILL AREA: (4) USE SUPPORT EQUIPMENT

3. MANIPULATE CONTROLS (NUMBER AND TYPE)

   A. SINGLE CONTROL - SINGLE TYPE; E.G., ON/OFF SWITCH, VALVE.

6. OPERATE SUPPORT EQUIPMENT (TYPE OPERATION)

   C. POWERED OPERATION - REQUIRES CONNECTION; E.G., NC-8, MOBILE AIR CONDITIONING UNIT, CRANE, AIR BOWSER.

SKILL AREA: (5) USE TEST EQUIPMENT

1. SELECT TEST EQUIPMENT (REQUIREMENT)

   C. REQUIRES MORE THAN ONE ITEM OF TEST EQUIPMENT - EACH SERVES MORE THAN ONE PURPOSE, E.G., MULTIMETER, TEST SET, AND OSCILLOSCOPE.

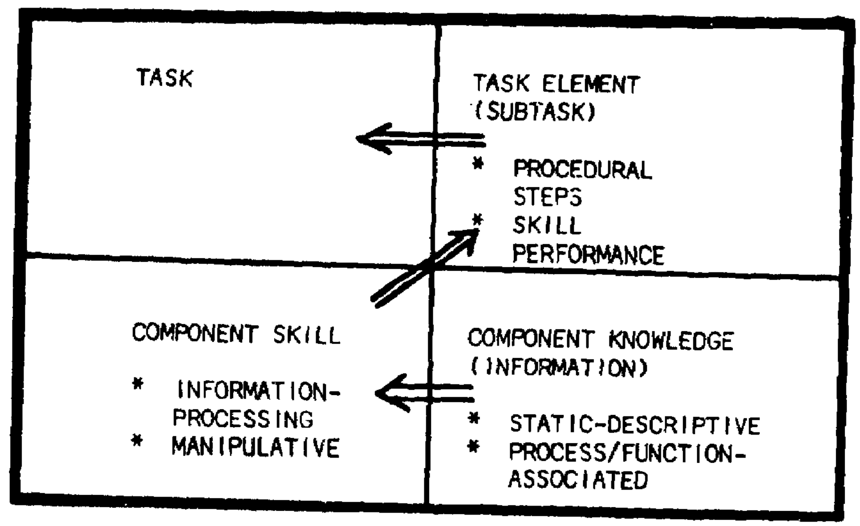
4. OPERATE TEST EQUIPMENT

   C. MUST BE CONNECTED TO SYSTEM AND REQUIRES MANUAL STEP-BY-STEP PROCEDURES TO OBTAIN READINGS.

FIGURE 3. SAMPLE SKILL-AREA CITATIONS PRINTED OUT IN NARRATIVE (AVIATION ELECTRICIAN'S MATE JOB/TASK INVENTORY

INVENTORIES

INVENTORIES DOCUMENTS (TECH)



INVENTORIES
TEXTS

DOCUMENTS (TECH)
TEXTS (TRAINING)
CATALOGS
INSTRUCTIONS/REGS
INVENTORIES

PROCESS/FUNCTION-ASSOCIATED
(STRUCTURED/ORGANIZED)

LAWS, FORMULAE, AXIOMS, HYPOTHESES, CONCEPTS, FUNCTIONS, TABLES, CHECKLISTS, LOGIC TREES, FLOW DIAGRAMS, SCHEMATICS, DECISION MATRICES, WORKING DRAWINGS

STATIC-DESCRIPTIVE

DATES, DIMENSIONS, STOCK NUMBERS, NOMENCLATURE, SHAPES, COLORS, IDENTIFICATION MARKS, CONTOUR, CONFIGURATIONS, NAMES, TIME, PLACES, CHARACTERISTICS, INDEXES, INVENTORIES, CATALOGS

FIGURE 4. RELATIONSHIP OF INFORMATION BASE TO TASK/SKILL PERFORMANCE

SKILL DATA LAYOUT — SKILL: SOLDERING (MANIPULATIVE)

ACTION: SOLDER

<u>GENERIC OBJECT(S)</u>

SHEET METAL(S)
CASTINGS
ELECTRICAL/ELECTRONIC CONNECTIONS
HAND-WIRED CHASSIS COMPONENTS
PRINTED CIRCUIT COMPONENTS
DISPARATE METALS

<u>APPLICATIONS</u>

FASTENING:

SHEET METAL TO SHEET METAL
CASTINGS TO CASTINGS
WIRES TO CONNECTORS (CLIPS, TABS, ETC)

MICROMINIATURE CIRCUIT ASSEMBLY/DISASSEMBLY

<u>METALS INVOLVED</u>

GALVANIZED IRON SHEET
BRASS, COPPER, ALUMINUM SHEET
BRASS, COPPER, BRONZE, ALUMINUM CASTING
SILVER, PEWTER, POTMETAL CASTING
COPPER, ALUMINUM WIRE

<u>TOOLS</u>

SOLDERING IRON, ELECTRIC:

LARGE HIGH-WATTAGE (SHEET METAL, DUCTING, ROOFING) GUN-TYPE, 100-300 WATT AC
PENCIL, 37-40 WATT AC

HEAT BLOWER, ELECTRIC, 2000 WATT AC
TORCH, PROPANE/BUTANE
DE-SOLDERING KIT (SOLDER SUCKER, IRON)

<u>SUPPORT MATERIALS</u>

SOLDER:

TIN, LEAD, ZINC COMBINATIONS (OR ALLOYS):
60/40, 50/50, 40/60
ACID-CORE, ROSIN-CORE
SILFOSS (SILVER SOLDER)
ALUMINUM

FLUX, SOLDERING PASTE, MURIATIC ACID
FILE, ABRASIVE CLOTH, SANDPAPER
BRUSH, CLEANING CLOTHS, PAPER TOWELS

<u>SUPPORT EQUIPMENT</u>

HEAT SINK
CLAMPS ("C" TYPE, SPRING LOADED, VISE GRIP)
MAGNIFYING GLASS
ASSEMBLY JIG
INSULATED IRON/GUN REST
HEAVY-GAUGE EXTENSION CORD
ABC-TYPE FIRE EXTINGUISHER

<u>TEST EQUIPMENT</u>

GENERAL PURPOSE (MULTIMETER ETC.)

<u>REFERENCES</u>

NAVEDTRA 1034-D, RTM-AVIATION ELECTRICIAN'S MATE 1 & 2
MODULE 8, BASIC ELECTRICITY & ELECTRONICS COURSE
TRAINEE GUIDE, AVIATION ELECTRICIAN'S MATE "A" SCHOOL COI.

FIGURE 5. SKILL-SUPPORTING INFORMATION PACKAGE FOR THE TRAINING PROGRAM DEVELOPER

MODEL COMPONENT WORK BEHAVIORS (PROCEDURES)

CHECK APPROPRIATE REFERENCE FOR SPECS/REQUIREMENTS OF SOLDERING TASK.

SELECT TOOLS, SUPPORT MATERIALS, SUPPORT EQUIPMENT, SOLDER (BY DETERMINING METALLURGY REQUIREMENTS FROM REFERENCE) (APPLICATION OF COGNITIVE).

PREPARE WORK ENVIRONMENT (REFERENCE TO JOB DATA, SAFETY REQUIREMENTS, ETC.)

CONNECT/WARM UP IRONS.

CLEAN JOINGS/PREPARE SURFACES FOR SOLDERING, COAT WITH FLUX

SET UP WORK ON BENCH, IN JIG, OR IN PLACE IN MAJOR ASSEMBLY.

APPLY CLAMPS, HEAT SINKS AS APPROPRIATE.

MANIPULATE SOLDERING TOOL, SOLDER TO PRODUCE SOLDERED SEAM/JOINT/CONNECTION.

INSPECT SOLDERED SEAM, JOINT, CONNECTION (USING OBSERVATIONS SPECIFIED IN REFERENCES).

EVALUATE, TEST COMPLETED TASK PRODUCT (USE TEST EQUIPMENT TO CHECK CONTINUITY THROUGH SOLDERED CONNECTION, APPLY BENDING OR OTHER STRENGTH TEST, APPLY NEATNESS, QUANTITY-OF-SOLDER OR ANY OTHER SPECIFIED STANDARD OF MEASUREMENT OR EVALUATION).

REMEDIATE, IF NECESSARY. RE-DO STEPS PROVEN ERRONEOUS, REJECT ITEM, DO ENTIRE TASK OVER, ETC.).

FIGURE 5 (CONTINUED). SKILL-SUPPORTING INFORMATION PACKAGE FOR THE TRAINING PROGRAM DEVELOPER

THE JOB/TASK ANALYSIS/SKILLS AND KNOWLEDGE MARRIAGE (PART II)

T. M. Ansbro and W. A. Hayes
CNET/TAEG NEPDIS Development Team, CNET HQ. NAS, Pensacola

This paper (Part II) illustrates input and eventual employment of model "tasks, skills and knowledges" in the front-end job/task analysis subsystem of NEPDIS (Naval Enlisted Professional Development Information System). The four-part matrix displayed in Part I reflects a hierarchy with the task at the top and the associated knowledge elements at the bottom. The task-to-task element-to-component skill-to-component knowledge continuum provides an audit trail for the use of the training program or curriculum developer, whether the matrix is added to the existing master job/task inventory or provided as an ancillary data bank for specialized use.

The principal goal is to set up a functioning audit trail (to justify a body of job-related technical information as actually component to or clearly underlying task performance). A secondary goal is to set up an occupational-field data bank and computerized retrieval methodology to support this aim. The outputs of front-end job/task/skill analysis can then be used both to describe (even construct) jobs/billets and the tasks performed by their incumbents, and to describe the skills and knowledge requirements for job incumbency, certification, advancement, and associated training.

This paper (Part II) illustrates the input and eventual employment of model "tasks, skills and knowledges" in the front-end job/task analysis subsystem of NEPDIS (Naval Enlisted Professional Development Information System). The principal goal is to justify a body of job-related technical information as actually component to or clearly underlying task performance; a secondary goal is to set up an occupational-field data bank and computerized retrieval methodology to support this aim.

The initial input into the model (Figure 1) is a comprehensive inventory of job/task statements consisting of specific actions to be performed on (or with) actual, real-world (not generic) task objects with conditions, standards and supporting descriptive factors. It cannot be over-emphasized that until all the above data as a minimum have been collected, no total and thorough analysis can be completed. The term "front-end analysis" as used in this paper is intended to portray an analysis of all data collected at the front end of any particular occupational data acquisition and analysis project. It is not applicable to any system which collects only a portion of the required data, conducts a portion of the analysis, and then returns to the same (or additional) sources for another round of data collection.

Once the data have been collected, a computerized analysis can sort tasks into skill levels, pay grades, or any other desired distribution based on the inherent complexity of the supporting descriptive factors (Figure 2). This first computer sort produces a crucial product by selecting from the total inventory those tasks most appropriate for assignment to a particular group of job incumbents (Figure 3). The initial list of selected tasks then serves as an input into other computer algorithms which produce lists of billet-specific tasks and a list of rating-specific skills (Figure 4). The skills are then further analyzed to determine knowledge requirements.

The billet-specific tasks can serve useful roles in determination of manpower requirements, assignment of personnel, identification of specific training requirements, advancement in rate, and certification and evaluation of workers' performance on the job. Rating-specific skills serve as the standard to which all members of a rating (regardless of billet to which assigned) must be able to function in order to perform the tasks of a specific pay grade, skill level, or other distribution structure. Knowledge requirements serve to support absolute standards for advancement in rating and for certification in work qualifications.

Once skills and knowledges have been identified and extracted from the tasks, the selection of training sites and methods becomes the area of primary concern (Figure 5). Skills and knowledges can be prioritized for training; and those with highest priority would be assigned to a schoolhouse setting (for instance a class "A" school). The remainder would appropriately be assigned to correspondence courses and other self-study modes. In most cases the training manuals used as the basic text for schoolhouse courses would also be used for follow-on self-study and correspondence courses.

Tasks, like skills and knowledges, can be prioritized for training. Those with high priority would be assigned to schoolhouse settings (for instance, class "C" and class "F" schools), while those of low priority would be assigned to formal on-the-job training.

A total training continuum based on the above assignment to the various training settings would ensure that a worker reporting to a new assignment would have completed training for the skill and knowledge requirements established as necessary to perform in his particular pay grade. It would also ensure that he had been trained to perform tasks peculiar to his particular billet. He would report to his work site already prepared to perform useful work and after a brief period of formal on-the-job training, would be fully billet-qualified and able to perform all tasks assigned his billet.

Job certification and advancement are other areas which would also be served by an automated analysis of occupational data (Figure 6). A task-specific billet description for a worker's next billet and the three elements of training--tasks, skills, and knowledges are the input items for the algorithm. A worker's ability to perform tasks and skills would be evaluated by hands-on performance tests and his possession of adequate knowledge requirements would be evaluated by paper and pencil tests. Once he had demonstrated proficiency in the three elements required for a new billet, he would become eligible for advancement. He could then advance into the new billet and begin functioning as a fully qualified and certified incumbent; upon meeting other requirements such as military leadership, time-in-rate, etc. he could automatically be advanced to the next pay grade commensurate with the billet. The Service as well as the individual would receive fair and equitable compensation.

RATING = AE   PACKAGE = 0001   TASK = 0020   DUTY SUBCATEGORY = 01

TASK STATEMENT = PERFORM MAINTENANCE CHECK ON ROTATING BEACON ASSEMBLY 44126

PLATFORM = P-3 A/B (APBC)
SYSTEM   = LIGHTING SYSTEM (44000)
EQUIPMENT = EXTERIOR LIGHTING (44100)
COMPONENT = ROTATING BEACON ASSEMBLY (44126)

COMPLEXITY = 1.67

MAJOR ACTION CATEGORY = MAINTAIN
DUTY SUBCATEGORY (01) = CHECK/TEST/INSPECT
TASK ACTION (PMC)   = PERFORM MAINTENANCE CHECK

CUE..........................REPAIR COMPLETE
STANDARD.....................IAW REFERENCE PUBLICATION
REFERENCE....................NA-01-75PAA-2-12
TOOL.........................COMMON HAND TOOLS, SPECIAL HAND TOOLS
SUPPORT MATERIAL.............FUSE/SWITCH
SUPPORT EQUIPMENT............POWER UNIT 28V 400HZ
TEST EQUIPMENT...............RPM GAUGE, AMMETER, MULTIMETER

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
GENERAL	2	1	1	3	1	1	1	1	2	0	0	0	0	0
DUTY SUB 01	2	3	2	0	1	3	1	0	0	1	0	0	0	0
SKILL 1 (REFERENCE)	2	1	1	1	1	0	0	0	0	0	0	0	0	0
SKILL 2 (TOOL)	3	3	2	1	0	0	0	0	0	0	0	0	0	0
SKILL 3 (SUPPORT MATRL)	2	0	0	0	0	0	0	0	0	0	0	0	0	0
SKILL 4 (SUPPORT EQUIP)	1	1	1	1	0	3	0	0	0	0	0	0	0	0
SKILL 5 (TEST EQUIP)	3	2	2	3	2	2	2	1	3	0	0	0	0	0

Figure [illegible]. SAMPLE TASK FROM JOB TASK INVENTORY

RATIONALIZATION OF JTI DISTRIBUTION INTO SKILL LEVEL/PAY GRADE GROUPS

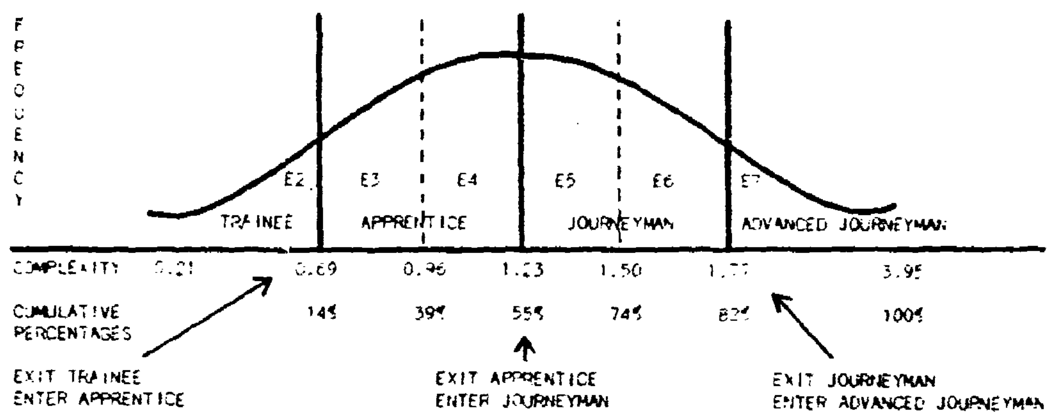


Figure 2. TASK DISTRIBUTION HISTOGRAM

DISTRIBUTION OF JOB TASK INVENTORY INTO SKILL LEVEL/PAY GRADE GROUPS

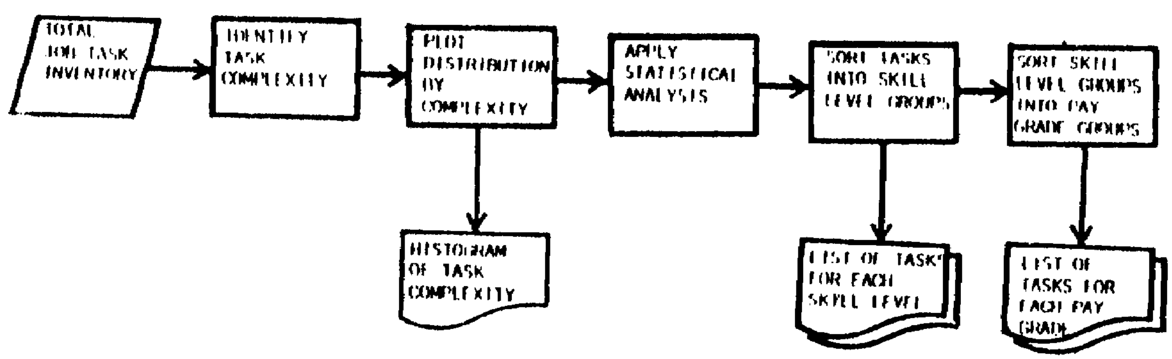


Figure 5. FLOW CHART OF SKILL LEVEL/PAY GRADE SHRED-OUT

IDENTIFICATION OF BILLET-SPECIFIC TASKS, RATING-SPECIFIC SKILLS AND RATING-SPECIFIC KNOWLEDGE REQUIREMENTS

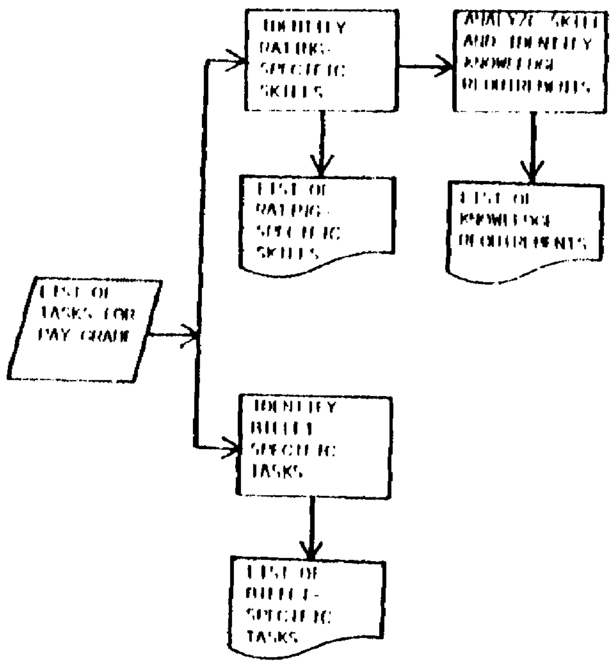


Figure 4. FLOWCHART, IDENTIFICATION OF TRAINING REQUIREMENTS

ASSIGNMENT OF TASKS, SKILLS AND KNOWLEDGES TO TRAINING SETTINGS

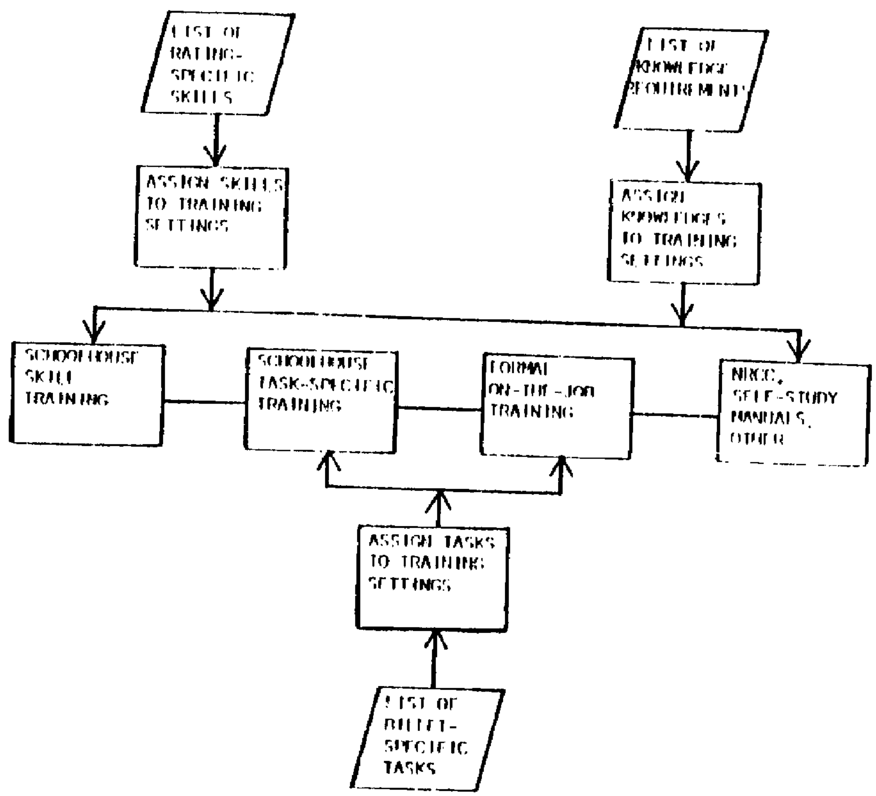


Figure 5. FLOWCHART, MATCHING TRAINING REQUIREMENTS WITH TRAINING SETTINGS

JOB CERTIFICATION AND ADVANCEMENT

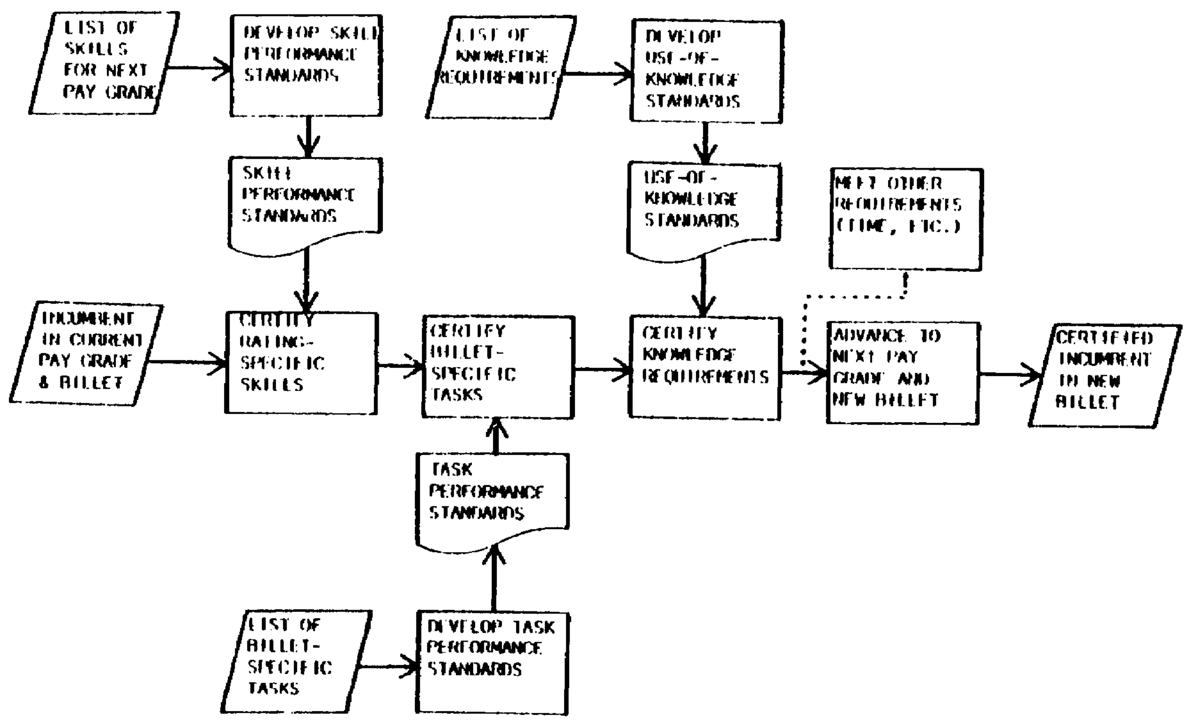


Figure 6. FLOWCHART, CERTIFICATION AND ADVANCEMENT

DEVELOPMENT OF THE GENERAL WORK INVENTORY

Rodger D. Ballentine, Maj
Air Force Institute of Technology

J. W. Cunningham, Professor of Psychology
North Carolina State University

AD P001288

The services have a sophisticated task-based job analysis system which provides invaluable training and job descriptive information within occupational areas. Increasingly, however, requests are made for the comparison of related work functions or variables (e.g., skills) across occupational areas. A practical quantitative measure for comparing work activities in different career fields would facilitate broad description, comparison, and classification of such occupational information. This paper outlines a study to develop a structured questionnaire consisting of work activity/condition descriptors applicable to the entire occupational spectrum. In addition to outlining the potential applications of such a system, we will describe plans to rate a sample of enlisted Air Force jobs, assess instrument reliability and validity, and group jobs based on their work dimension profiles.

INTRODUCTION TO WORK STUDY

Historically, the idea of man's work has religious, ethical, social, and economic implications. Work is either a burdensome necessity, a means to an end, or a creative and valued act of man (Prien & Ronan, 1971). In this discussion, work is viewed as a process whereby one exerts effort to transform various inputs into prescribed outcomes. Since the turn of the century, job analysis has been synonymous with work study, as explained by McCormick and Tiffin (1974): "Job analysis can be considered as embracing the collection and analysis of any type of job-related information, by any method, for any purpose. Perhaps it can be defined more generally as the study of human work" (p. 49).

The field of job analysis includes both qualitative (conventional) and quantitative (structured) methods for collecting worker attribute and work performance information. The conventional approach is useful for describing specific jobs; however, such information is not generalizable. On the other hand, the structured approach provides a consistent framework for data collection and therefore the ability to compare and numerically classify the units of analysis. Frequently confusion in this area of investigation results from inconsistently defined and applied terminology. In an effort to reduce further confusion and assist the reader, several terms pertinent to this discussion are defined below.

A work element is a description of various kinds of work activities or conditions on which positions or jobs can be rated. Typically, such elements are general enough to be applicable to a wide variety of jobs.

A work dimension is a statistically derived construct representing work elements which commonly occur together in positions/jobs.

A job is a group of positions in which major tasks are similar enough to justify a single analysis within or across organizations.

A job family/cluster is a group of jobs or occupations which have common characteristics.

PREVIOUS PAGE IS BLANK

The quantitative approach promotes more objective and systematic investigation of the relationship between work units. Commonly tasks, duties, and work elements/dimensions are used to analyze positions, jobs, or occupations. Molecular analysis of task and duty similarity between positions or jobs is typically done within the same organization, whereas more molar analysis of work element/dimension similarity between jobs and occupations is typically done across organizations. The relationship between these units of job analysis form a hierarchy for the study of work (Pearlman, 1980). Individuals performing tasks represent positions, positions containing similar tasks are grouped to form jobs, and jobs with common work elements/dimensions are grouped to form job families. A pictorial representation of these levels of work and descriptive measures, adapted from Pearlman (1980), is presented in Figure 1.

QUANTITATIVE WORK TAXONOMIC RESEARCH

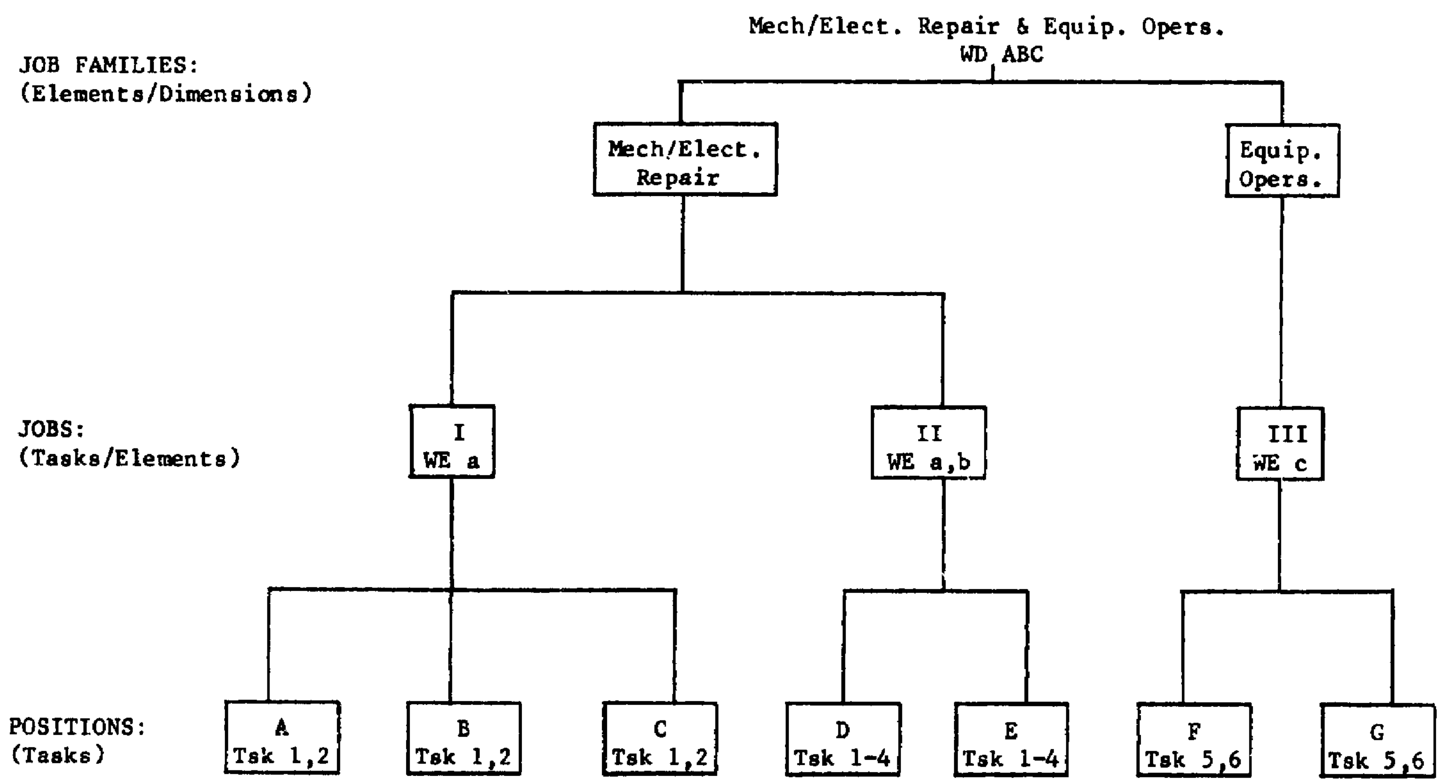
Prien and Ronan (1971) drew two conclusions from a review of job analytic research. First, they found that few studies provide basic and generally useful information about work, workers, and occupations. Second, data from various job taxonomic efforts were not comparable because of procedural and measurement differences. Jones and DeCoths (1969) conducted a national survey of firms about their use of job analysis and attributed management dissatisfaction with these programs to a lack of standard quantitative techniques for gathering, recording, and presenting job information. Pearlman (1980) echoed the importance of systematic job analytic procedures in the development of a taxonomy of work performance:

> The development of strategies for classifying and grouping jobs in some systematic fashion thus appears to be an essential step in the effort to devise a unified taxonomy of work performance, that is, one that addresses the relevant characteristics of both people and jobs. (p.3)

From both the theoretical and applied standpoint, standard quantitative work measurement procedures are important in the study of work.

Systematic procedures for studying jobs are available from quantitative job analytic research. These methods most often employ structured job questionnaires to gather information about the work accomplished in jobs within specified occupational areas. Probably the best example of this methodology is the task-inventory approach developed by the USAF and other services (Christal, 1974). Overall, the structured job analysis questionnaire has proven to be an economical and reliable tool for collecting job information. Taxonomic research to define common work denominators suggest new uses for this data collection technology.

Paralleling these improvements in data collection are conceptual advances promoting theoretical and practical interest in work taxonomic research (Cunningham, 1974; McCormick, 1979; McKinlay, 1976; Pearlman, 1980). Cunningham (1974) called the required technology for taxonomic definition and measurement of work "Ergometrics," and defined it as "the application of psychometric principles and procedures to the study of human work" (p. 7). Both McCormick and Cunningham have identified work activity/condition descriptors to serve as common denominators in job description and classification. These descriptive statements, applicable to jobs throughout the occupational spectrum, are promising variables for a comprehensive work taxonomic system. Moreover, job elements can be linked to defined human



<u>Tasks</u> (Tsk)
1. Diagnose engine problem
2. Service/adjust engine
3. Isolate malfunction
4. Remove or install wires
5. Start-up equipment
6. Control equipment

<u>Work Elements</u> (WE)
a. Mechanical repair
b. Electrical repair
c. Equipment operations

<u>Work Dimensions</u> (WD)
ABC. Mech/Elect repair & equip. opers.

Figure 1. An example of work analysis within an organization. Adapted from Pearlman (1980).

attributes for which there are tests (McCormick, Jeanneret, & Mecham, 1972; Pass & Cunningham, 1975), and can be used to cluster jobs with similar activity or attribute requirements (Pass & Cunningham, 1978; Shaw, DeNisi, & McCormick, 1977).

It is our thesis that the required tools for a comprehensive work description and classification system are psychometrically based job activity questionnaires (Cunningham, 1974). Two such questionnaires have been constructed with consideration of the "job-oriented" and "worker-oriented" dichotomy in work activity descriptors proposed by McCormick (McCormick, Cunningham & Gordon, 1967). The Position Analysis Questionnaire (PAQ) is primarily made up of worker-oriented activity statements (McCormick et al., 1972), while the Occupation Analysis Inventory (OAI) is both job- and worker-oriented (Cunningham, Tuttle, Floyd, & Bates, 1974). Both instruments portray work through an information-processing model with similar item categories. These inventories are representative of the state of the art in collection of quantified job activity information, and their descriptive taxonomies reflect job- and worker-oriented activities common to the world of work. However, widespread data collection with both of these instruments is somewhat limited by their complexity and the demands they place on the rater.

PROPOSED STUDY

Existing job activity inventories are often impractical because trained personnel or highly educated raters are required to collect job data with them. This study has two primary goals. First, we plan to develop a generally applicable structured work questionnaire which is simple enough to be completed by job incumbents. This instrument should be brief, clear to the typical job holder, and present a straightforward job rating task. Secondly, we hope to demonstrate how the questionnaire can be used to define activity/condition dimensions of jobs and how these dimensions can be used to meaningfully classify jobs. A practical questionnaire would not only facilitate the collection of quantitative information about jobs, but would also promote the use of a common language between employer and employee. A description and classification system based on such an instrument could find use in both personnel management and employment counseling.

In the proposed study, we plan to apply the previous OAI research to the following objectives:

a. Develop a set of work descriptors based on factor analyses of the original OAI items by Boese and Cunningham (1976).

b. Incorporate the resulting work variables into a structured questionnaire which can be administered to job incumbents. This instrument will be referred to as the Ceneral Work Inventory (GWI).

c. Apply the GWI to a sample of Air Force jobs performed by skilled personnel. Different rater groups will analyze the same and different jobs. The resulting ratings will be used to determine item reliabilities within groups and rater agreement across groups.

d. Determine the factor structure of the GWI elements and investigate the construct validity of the resulting GWI dimensions. Dimension validity will require evidence of factor stability and relevance to human attribute and job content criteria. For example, relationships between the GWI dimensions and worker aptitudes will be determined by regressing the GWI factor scores and aptitude-requirement estimates for a sample of jobs against their mean incumbent aptitude test scores. (Established aptitude-requirement indices for jobs, such as relative learning load or cutoff scores, might also serve as criteria in the regression analyses.) In addition, it will be determined if existing Air Force job categories (i.e., content groups and groups based on the principal aptitude requirement) are significantly discriminable in terms of their mean GWI factor scores and aptitude-requirement estimates.

e. Apply the GWI dimensions to the development of a job taxonomy by grouping the sample jobs on their factor-score profiles. The obtained cluster structure will be checked for stability and meaningfulness. The Air Force classification scheme and overall job-similarity groupings will be used as criteria for comparison to the GWI-based job taxonomy. Additionally, if the GWI dimensions are relevant to aptitude requirements, the mean aptitude test scores of job incumbents should differ significantly across clusters.

The GWI is designed as a short version of the OAI which can be administered to job incumbents and experts. "Man is viewed as an information processing system which transforms information input into prescribed outcomes" (Cunningham et al., 1974, p. 8). This framework is suited to the process view of work (i.e., input-throughout-output) described earlier and relates to existing inventories as presented in Figure 2. The GWI elements are generated primarily from the OAI work dimensions, which represent types of tasks and conditions commonly occurring in the world of work. These elements should be suitable for job description and relevant to measured human attributes. The dimensions derived from GWI elements will cover broader aspects of related work which can be used to profile jobs for description, comparison, and classification purposes.

The target population in this study consists of the jobs performed by skilled (i.e., 5-level) enlisted personnel in the USAF. The Air Force Military Personnel Classification System groups positions in which related work is performed into specialties. (Jobs and specialties will be used interchangeably in this discussion.) The underlying principle of specialty formation is that the positions included have similar work requirements and therefore require similar abilities. The specialty will be the primary unit of analysis in this study. A sample of at least 200 5-level jobs (specialties) will be selected to represent the existing classification scheme. This sample should provide a reasonable representation of the ability and other work requirements in the target population.

Participants in this study will perform one of the following tasks: rate a position or job with the GWI, judge the pairwise similarity within a job subsample or rate the relevance of Armed Services Vocational Aptitude Battery (ASVAB) composite areas to GWI elements. Four different groups of personnel will rate enlisted 5-level specialties: Job incumbents (INC), their supervisors (SUP), occupational analysis specialtists (OAS), and promotion test development personnel (PTD). Ratings by non-INC groups will be obtained to determine the reliabilities of different rater sources and agreement among these sources. Senior promotion test reviewers and occupational analysis personnel from the Occupational Measurement Center and career field functional representatives from the Manpower and Personnel Center will each judge the pairwise similarities of approximately 25 jobs. (All non-INC/SUP job raters and similarity judges will review formal job descriptions and relevant occupational analysis information when completing their respective rating task.) The pairwise job similarity ratings will be used to create a job similarity matrix as a basis for grouping jobs. Several ASVAB test development psychologists from the Human Resources Laboratory and personnel

Position Analysis
Questionnaire (PAQ)

Information Input

Mediational Processes

Work Output

Interpersonal Activities

Work Situation &
Job Context

Miscellaneous Aspects

Occupation Analysis
Inventory (OAI)

Information Recieved
--Information Content
--Sensory Channel

Mental Activities

Work Behavior
--Physical
--Representational
--Interpersonal

Work Goals

Work Context

General Work Inventory (GWI)

What is Worked on

Behavioral Processes

What is Accomplished

Contextual Conditions

Figure 2. Work analysis frameworks.

psychology graduate students and staff from North Carolina State University will rate the relevance of major ASVAB areas to GWI elements. The mean relevance ratings for each element will be used as weights to derive aptitude-requirement estimates for jobs, according to McCormick's job component validity procedure (McCormick, Cunningham, & Thornton, 1967).

The process of collecting GWI job ratings is relatively straightforward. Questionnaires will be mailed to INC and their SUP, and ratings by OAS and PTD will be controlled by a liaison at the Occupational Measurement Center. Several INC in each job will rate their positions with or without the assistance of their SUP. A subset of 40-50 jobs, in career ladders undergoing Staff Sergeant promotion test revision during the administration period, will be rated by all groups. For these commonly rated jobs, INC and their SUP will independently rate the INC's position; whereas in all the other 150-160 jobs INC and SUP may work together to rate the INC's position. In order to collect three ratings for each job from non-INC/SUP groups, we will ask each OAS to rate about four of the common jobs and each PTD to rate the 5-level job for the career ladder in which they are developing the promotion test.

The following analyses will be performed:

a. Interrater item reliability analyses. These analyses will be carried out within each of the four rater groups (INC, SUP, OAS, and PTD), and ratings will be compared for agreement across the four groups.

b. Computation of GWI aptitude-requirement estimates for jobs, following McCormick's job component procedure. These estimates will be derived by combining job ratings on the GWI elements with those elements' aptitude-requirement weights.

c. Factor analyses of the GWI elements. These will include factor analyses of sections (or groups) of GWI elements, as well as an overall analysis of the entire set of elements. A factorial stability analysis will involve: (1) dividing the job sample into two equivalent subsamples, (2) performing independent factor analyses with both subsamples, and (3) comparing the results for replication across the subsamples.

d. Cluster analyses of jobs. The jobs in the sample will be cluster analyzed on two bases: (1) similarities among their GWI factor-score profiles and (2) analysts' pairwise job similarity judgments. The clusters derived from these two data bases will be compared for agreement with each other and with the Air Force classification scheme.

e. Several construct validation analyses will be carried out, including: (1) regression analyses of jobs' GWI factor scores and ability-requirement estimates against the ability test scores of job incumbents (and possibly against the established test cutoff and relative learning load scores for the jobs); (2) analyses to determine the discriminability of GWI-derived job clusters in terms of the ability test scores of incumbents in the jobs comprising the clusters; (3) analyses to determine the discriminability of existing Air Force job categories (both content and aptitude groupings) in terms of the GWI factor scores and ability-requirement estimates of the jobs comprising those categories; and (4) the cluster comparisons mentioned in paragraph d above.

WORK CLASSIFICATION SYSTEM UTILITY

Successful performance of the GWI in this study would support continued research of this general approach in Air Force work analysis. An effective occupational information system would cover a range of descriptive specificity, from task statements (applicable to restricted groups of positions), to more general work elements and dimensions (applicable to the broad occupational spectrum). The utility of the system would also depend on its ability to interrelate the different types of descriptive variables (e.g., tasks, work elements/dimensions, and human attributes). The task inventory would remain the major methodological component in such a system, supplemented by one or more instruments containing descriptors of a more general nature.

A quantitative work taxonomy would have numerous applications for describing, relating, and researching the characteristics of jobs and workers. The identification of work variables applicable to the entire occupational domain could provide a basis for describing and classifying a wide variety of jobs within the same system. The linkage of these work descriptors to measurable human attributes would permit determination of relationships between people and jobs. In a sense, this relational mapping could serve as a table of contents for understanding commonalities in the work activities and human requirements of jobs and job families. Such a system could facilitate a variety of human resource development efforts. For example, applications might be found in such areas as occupational exploration and guidance, recruitment, job transfer, career development, training, test development, and selection/assignment.

REFERENCES

Boese, R. R., & Cunningham, J. W. Systematically derived dimensions of human work. JSAS Catalog of Selected Documents in Psychology, 1976, 6, 57. (Ms. No. 1270).

Christal, R. E. The United States Air Force occupational research project (AFHRL-TR-73-75). Lackland Air Force Base, Tex.: Air Force Human Resources Laboratory, Occupational Research Division, January 1974. (NTIS No. AD 774 574)

Cunningham, J. W. "Ergometrics": A systematic approach to some educational problems. JSAS Catalog of Selected Documents in Psychology, 1974, 4, 144-145. (Ms. No. 804)

Cunningham, J. W., Tuttle, T. C., Floyd, J. R., & Bates, J. A. The development of the Occupation Analysis Inventory: An "ergometric" approach to an educational problem. JSAS Catalog of Selected Documents in Psychology, 1974, 4, 144. (Ms. No. 803)

Jones, J. J., Jr. & DeCoths, T. A. Job analysis: National survey findings. Personnel Journal, 1969, 49, 805-809.

McCormick, E. J. Job analysis: Methods and applications. New York: AMACOM, 1979.

McCormick, E. J., Cunningham, J. W., & Gordon, G. G. Job dimensions based on factorial analyses of worker-oriented job variables. Personnel Psychology, 1967, 20, 417-430.

McCormick, E. J., Cunningham, J. W. & Thornton, G. C. The prediction of job requirements by a structured job analysis procedure. Personnel Psychology, 1967, 20, 431-440.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). Journal of Applied Psychology, 1972, 56, 347-368. (Monograph)

McCormick, E. J., & Tiffin, J. Industrial psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1974 (6th Edition).

McKinlay, B. Characteristics of jobs that are considered common: Review of literature and research. Columbus, Oh.: The Center for Vocational Education, 1976. (ERIC No. ED 141 638)

Pass, J. J., & Cunningham, J. W. A systematic procedure for estimating the human attribute requirements of occupations. JSAS Catalog of Selected Documents in Psychology, 1975, 5, 353. (Ms. No. 1151)

Pass, J. J., & Cunningham, J. W. Occupational clusters based on systematically derived work dimensions: Final report. JSAS Catalog of Selected Documents in Psychology, 1978, 8, 22. (Ms. No. 1661)

Pearlman, K. Job families: A review and discussion of their implications for personnel selection. Psychological Bulletin, 1980, 87, 1-28.

Prien, E. P., & Ronan, W. W. Job analysis: A review of research findings. Personnel Psychology, 1971, 24, 371-396.

Shaw, J. B., DeNisi, A. S., & McCormick, E. J. Cluster analysis of jobs based on a revised set of job dimensions from the PAQ. Department of Psychological Sciences, Report No. 3. West Lafayette, Ind.: Purdue University, April 1977. (NTIS No. AD A042 356)

Beaton, Albert E. & Barone, John L. Educational Testing Service, Princeton, New Jersey. (Thurs. P.M.)

The Usefulness of Selection Tests

In many selection situations, some predictor of success, such as a test, is used to discriminate between the applicants to be accepted and those to be rejected. Using an interpretation of Brogden (1946), this paper shows the effect of using a correlated variable on the mean criterion score of the selected group. The distribution of the means of all possible selections is discussed and a simple formula is provided for calculating the probability of achieving as high or higher mean using random selection as opposed to using a correlated variable. It is shown that even fairly small correlations result in selection decisions that are not often exceeded by chance. The Brogden interpretation and significance test is shown to extend simply to the case of several predictors.

THE USEFULNESS OF SELECTION TESTS

Albert E. Beaton and John L. Barone

Educational Testing Service

Introduction

In many admissions decisions a selection test is used to decide which members of an applicant pool will be selected for the available openings. The correlation coefficient between the selection test and the college's criterion of success, Grade Point Average, say, are often in the range of .40 to .55 for the applicants who are selected and then enroll. Such correlations are often squared and then interpreted as the proportion of variance explained by the predictor. Such a statement is true but is very difficult for an admissions officer to relate to what the effect is on his entering class. Some better interpretive scheme is needed.

The purpose of this paper is to show the effect of using a selection test on the average criterion score of the entering class. The correlation coefficient (times 100) is shown to be the percentage of improvement of using the selection test over what would happen on the average if the test were not used. Further, a simple formula is developed for approximating the probability that selection by chance would yield an entering class as able or abler than a class selected with the help of a valid predictor. It is further shown that, for reasonably sized colleges and with reasonable assumptions, selection using even small correlations will almost certainly result in an entering class that would earn higher criterion scores than would occur without the selection test.

1. The Effect of Different Selection Procedures

Let us assume that an admissions officer has the task of filling n available openings in a college from a pool of N applicants. We will call the selection ratio $\underline{f} = n/N$. We will assume throughout that his aim is to select the "best" applicants in the sense that they are the students who will perform best on some criterion of college success such as grade point average (GPA). The score on the criterion variable is not, of course, known at the time of admission decision and is never known for applicants who are not accepted and/or do not enroll.

Let us now number the applicants with an index i $(i=1,2,\ldots,N)$ and name the criterion variable y. The applicant pool is the population of interest; we will not assume that it is a sample from some larger population. The score of applicant i on the criterion variable y is y_i which is the score that the applicant would receive on the success criterion if the applicant were selected and did enroll. For convenience and without loss of generality, we will consider the values y_i to be in standard form, that is, to have an average value $\bar{y}$ of zero and a standard deviation S of one. We do not assume, at this point, that the distribution of the y_i is normal.

Let us first consider what would happen in some extreme selection situations. First, let us assume that the admissions officer is prescient, that is, that he knows exactly how each applicant would perform on y if admitted and enrolled. In this case, the admissions process would be straightforward; the n candidates who would receive the highest y_i would be accepted and in this way the average score on the criterion would be maximized. The decision rule is clear, except for tied y_i at the cut-off point which could be selected randomly. We will call the group selected this way the optimum group and the mean criterion score of that group will be called $\bar{y}_1$. Another extreme admissions situation is the case in which the admissions officer knows nothing about the applicant pool, at least nothing that is related to the success criterion. We can consider this case as knowing the applicants index number, i, and no more. In this case, the admissions officer might select applicants at random and hope for the best. There is a tiny probability that the resultant admittees would be the optimum group, but there is also a tiny probability of selecting the group with the lowest possible average score on y. Yet, with no useful predictive information available, there would be little else that could be done.

However, although we do not know what <u>will</u> happen if the applicants are selected at random, we do know everything that <u>could</u> happen. First, we know that there are precisely

$$C = \frac{N!}{n!(N-n)!} \tag{1}$$

different ways in which n students can be selected from a pool of N applicants. Let us number them with the index c $(c=1,2,\ldots,C)$. Each possible selection would result in a mean score for the criterion which we will label $\bar{y}_c$. We also know that the average of all possible $\bar{y}_c$ is

$$\text{ave}(\bar{y}_c) = \bar{y} = 0 \tag{2}$$

that is, not surprisingly, that the average of all samples of size n is the average of the applicant pool which is zero. We also know that the variance of the C possible $\bar{y}_c$ is

$$\text{var}(\bar{y}_c) = S^2\left(\frac{1}{n} - \frac{1}{N}\right) = S^2\left(\frac{1-f}{n}\right) = \frac{1-f}{n} \tag{3}$$

since the variance of the y_i is unity. Proofs of the average and variance are shown in Cochran [1977]. If the student selection were at random, then any of the C groups is as likely to occur as any other, and thus the average mean criterion score over a large number of random assignments would approach zero. We will label the average mean as $\bar{y}_0$.

Let us now consider a less extreme--and more realistic--situation in which the admissions officer has some imperfect information about the applicants. Let us assume, for example, that each applicant's dossier contains his score on an admissions test. We will call that test x and the score of applicant i on that test x_i ; x_i is known for all applicants. For convenience, we will also assume that the scores on x are in standard form with zero mean and unit variance. Let us also assume that, from experience, we know that the linear correlation of x with y is ρ . We will assume throughout that $\rho \geq 0$. One possible admissions rule is to select the n students with the highest scores on x . The question we wish to explore here is how much better will the admittees be if selected using x rather than selecting at random.

We do know something about what will happen if x is used for selection. First, we can partition the criterion scores into two parts

$$y_i = \rho x_i + e_i \tag{4}$$

where the first part ρx_i is the "predicted" value of y_i and e_i is a residual. Let us call the average y_i for the selected group $\bar{y}_\rho$. If we select the n applicants with the highest scores on x , then the value of

$$\bar{y}_\rho = \frac{\rho}{n} \Sigma_+ x_i + \frac{1}{n} \Sigma_+ e_i \tag{5}$$

where Σ_+ means summing the scores of the applicants who have the top n values of x . If the relationship between y and x is linear, that is, the conditional mean of y given x is ρx , then the term $\frac{1}{n}\Sigma_+ e_i$ will vanish and thus we may say

$$\bar{y}_\rho = \rho \bar{x}_+ \tag{6}$$

where $\bar{x}_+$ is the average of the top n applicants.

Comparison of the average criterion score of selected groups under optimum information, $\bar{y}_1$, no information, $\bar{y}_0$, and under information from some correlate, $\bar{y}_\rho$, leads to an interesting interpretation of the correlation coefficient which was first proposed by Brogden [1946]. Under the assumption that the distribution of y is the same as the distribution of x , then the average of the top n scores on y is the same as the top n scores on x , although the actual persons having those scores may be different. Under this assumption, therefore, $\bar{x}_+ = \bar{y}_1$, and

$$\rho = \frac{\bar{y}_\rho}{\bar{y}_1} \tag{7}$$

that is, the correlation coefficient (not squared) is the ratio of the mean of the group selected using x to the mean of the optimum group. Since the average mean of random selections is zero, 100ρ may be interpreted as the percent of possible improvement in selection attained by using the selection test over what would happen, on the average, using random or any arbitrary selection procedure that is independent of y .

Figure 1 shows the relationships among $\bar{y}_0$, $\bar{y}_\rho$, and $\bar{y}_1$ in the case where the top 10% of the applicants are to be selected. The abscissa represents scores on y and the ordinate is proportional to the number of applicants with a particular score. The normal distribution was selected for pictorial purposes. The shaded portion of the curve represents the top 10% of the applicants which has a cut-off score at +1.28. The average score of the top 10% is a +1.755 which is the mean of the optimum group, $\bar{y}_1$. The average random selection, $\bar{y}_0$, is zero. The line on the graph between $\bar{y}_0$ and $\bar{y}_1$ has tick marks to show where the mean of the selected group would be if the correlation with the selection test, ρ, were +.10,.20,...,.90. If $\rho = 0$, then the effect of the selection test would be the same as the $\bar{y}_0$ and if $\rho = 1$, the effect is the same as $\bar{y}_1$.

Insert Figure 1 about here

---

2. The Probability of Exceeding $\bar{y}_\rho$ by Chance

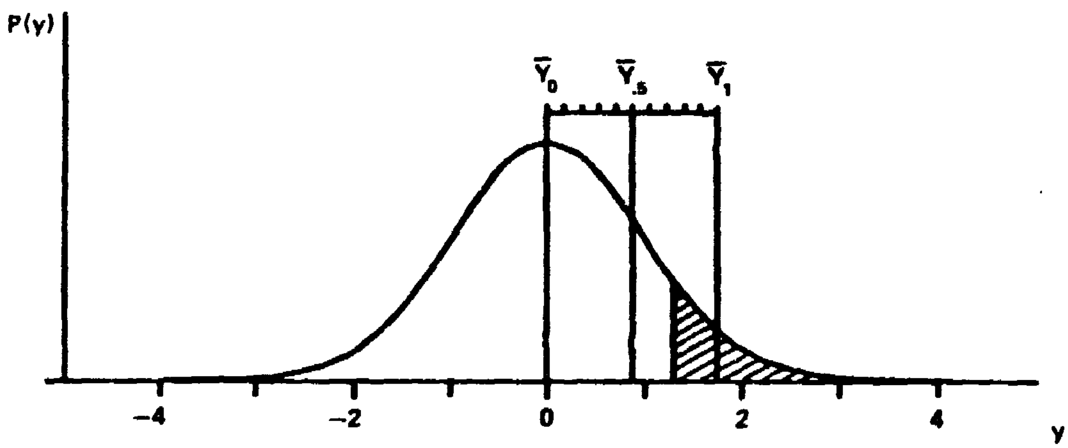
The improvement of the mean of the entering class using a small correlation, $\rho = .10$ or $.20$, say, may not, at first, seem worth the cost of the testing effort, but this is seldom so. The alternative would be to leave the selection to chance, since there is at least some probability that the chance selection would be better than the selection by a test. The question to be asked now is: what is the probability of actually selecting a sample by chance which results in an average y as high or higher than the mean of those selected by a test? The answer to this question is conceptually simple since all we would have to do is enumerate all C possible samples, count the number of samples which have means as high or higher than $\bar{y}_\rho$, and divide that number by C to find the proportion as high or higher than $\bar{y}_\rho$. However, this direct solution is not feasible since even for a small problem like selecting 10 or 90 out of a sample of 100 has a value of C is approximately 1.73×10^{13}. However, a reasonable approximation to that proportion is possible and used here.

The ability to compare random selection against selection by x gives an opportunity to show the importance of even small correlations in selecting

FIGURE 1

DISTRIBUTION OF APPLICANT POOL

$\bar{Y}_0$ IS THE AVERAGE OF ALL RANDOM SELECTIONS
$\bar{Y}_{.5}$ IS THE AVERAGE OF GROUP SELECTED WITH $\rho = .5$
$\bar{Y}_1$ IS THE AVERAGE OF THE OPTIMUM GROUP
$f = n/N = .1$
CROSSHATCH AREA REPRESENTS TOP 10% OF DISTRIBUTION



an entering class. Figure 2 shows the probability of doing as well or better by chance than by a selection test using the assumption that the distribution of y in the unselected population was normally distributed with zero mean and unit variance.

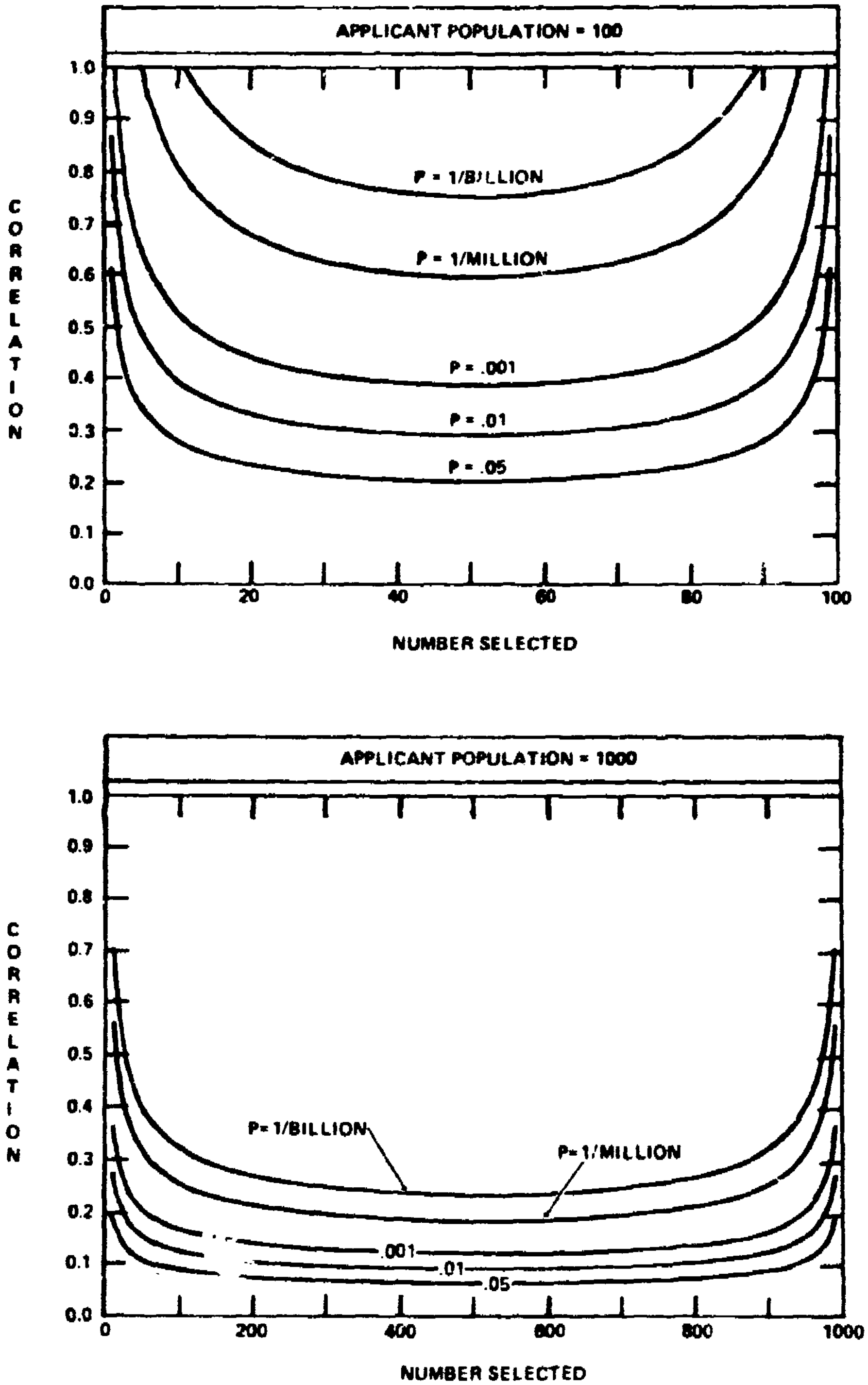
Probabilities are graphed for applicant pools of 100 and 1000, for different selection ratios, and for different values of ρ. We see that, in selecting 50 out of 100 applicants, a correlation of .30 is sufficiently high that chance selection would yield a higher mean than selection using the test not more than one time in a hundred. If the correlation is .40, the selected group may range from 10 to 90 and the probability that a group selected by chance will not excel the predicted group will not exceed one in a hundred. With an applicant pool of 1000, a correlation of .10 is large enough so that the probability of selecting a group of 500 with a larger mean by chance selection is less than .01 and a correlation of .20 is large enough so that the probability of selecting a group with a larger mean by chance is about 1/1000 or less. Thus, the use of a test with even a modest correlation will almost certainly result in a higher average of y.

Insert Figure 2 about here

Figure 3 shows why even such small correlations are useful in selection. Superimposed on Figure 1 is the approximate distribution of $\bar{y}_c$ for the case where the applicant pool is 100 and but 10 are to be selected. The mean of this distribution is zero and the standard deviation is $\sqrt{.09} = .3$. The distribution of $\bar{y}_c$ is very tall near its mean when compared

FIGURE 2

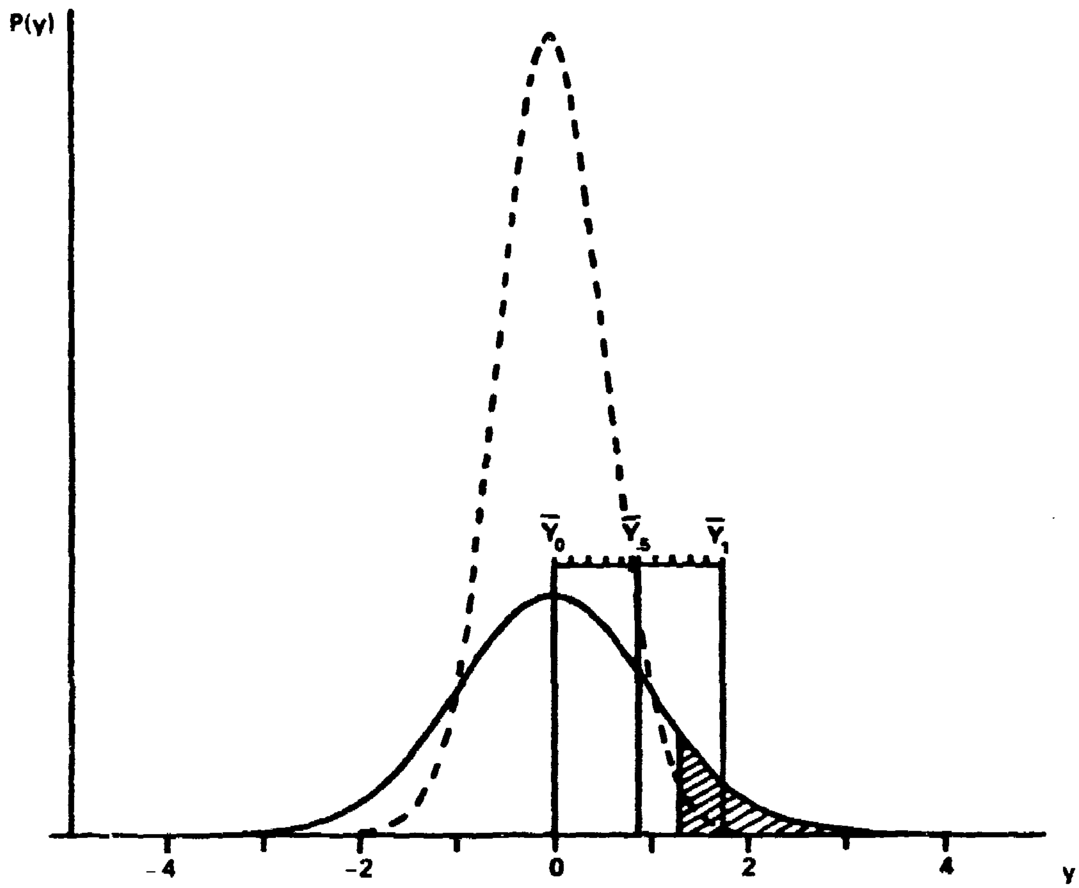
PROBABILITY OF $\tilde{Y}_c > \tilde{Y}_p$ FOR VARYING N,n and ρ



to the distribution of y_i which indicates that most of the $\bar{y}_c$ are close to the average and there are few large deviant values. It is true that the maximum value of $\bar{y}_c$ is $\bar{y}_1$, but the probability of this occurring by chance is approximately $.577 \times 10^{-13}$. It is to this distribution of $\bar{y}_c$ that the value $\bar{y}_\rho$ should be compared and it is clear that even for this very small problem the admissions officer is unlikely to improve on $\bar{y}_\rho$ by chance. For more realistic situations where the applicant pool is much larger, the distribution of $\bar{y}_c$ is even more spiked and thus the probability of exceeding $\bar{y}_\rho$ by chance is even less likely for many reasonable values of ρ .

Insert Figure 3 about here

FIGURE 3
DISTRIBUTION OF MEANS OF RANDOM SELECTIONS
N=100 n=10
DOTTED LINE REPRESENTS DISTRIBUTION OF ALL $\overline{Y}_c$



References

Brogden, Hubert E. On the interpretation of the correlation coefficient as a measure of predictive efficiency. Journal of Educational Psychology, 1946, 37, 65-76.

Cochran, William G. Sampling Techniques (third edition), New York: John Wiley & Sons, 1977.

AD P001290

CODAP - AN APPLICATION TO TRAINING
INFORMATION SYSTEM ANALYSIS

C. Derek Beel
Training Design Officer (CODAP)
Royal Naval School of Educational and Training Technology
HMS NELSON, Portsmouth P01 3HH, England

Summary

In October 1980 a Fleet Management Services Team began a study of the training management systems at three of the Royal Navy's apprentice training establishments. The aim was to formulate system parameters for a computer in each establishment to replace and enhance the method of gathering, processing and retrieving information on trainee performance and training management/assessment. For budgetary reasons the project had to be completed in 13 weeks. It was considered impossible to achieve this using conventional interview/analysis processes, and a system for gathering data by questionnaire for CODAP analysis was devised. The clustering program OVLGRP proved to be a very powerful tool for information system analysis and the method adopted was successful in defining the computer record and system requirements. The project demonstrated that valid results could be obtained from a complex questionnaire and indicated that the technique may be useful in management/organisational studies generally.

****

List of Annexes

A. Questionnaire Extract - Information Usage Section.
B. CODAP Processing Plan.
C. Training Management Computer Record - CODAP Extract
D. Information Usage Profile - HMS FISGARD Group 6.
E. Information/Job Time Dossier - HMS FISGARD Group 6.

Reference

Pass, J. J. & Robertson, D.W. Methods to evaluate scales and sample size for stable task inventory information. NPRDC, San Diego, May 1980.

INTRODUCTION

1. The Royal Navy has used CODAP for job analysis as an input to training design, Branch structuring, opinion and attitude surveys since 1973. Recently it has been "marketed" as a useful and powerful tool for the analysis of any manpower-based data and for any purpose. The aim of this short presentation is to describe the use of CODAP in the analysis of an information system, as an aid to a management services survey.

DISCUSSION

2. In October 1980, a Fleet Management Services Team was tasked with producing the system requirements for training management computers in 3 of the Navy's apprentice training establishments. The survey was concerned with the information generated, stored and used in the management of training from individual details (name, date of birth, religion etc.) through individual progress reports, class and course results to statistical analysis of tests. The main aims of the survey were:

a. To identify the data to be held on the computer to satisfy the needs of training management.

b. To develop a model for the proposed ADP system.

c. To identify duplication of information input, processing and output to point to organisational improvement (and possible staff savings).

d. To identify mis-matches between the information system and organisational structure.

There was one other aim:

e. To test the questionnaire and CODAP to establish whether it should be used for future similar projects.

3. For financial budgetary reasons the system requirements had to be defined in the short space of 13 weeks from the start of the project. At discussions between the survey team and my organisation, the RN School of Educational and Training Technology (RNSETT) it was agreed that the only feasible method of achieving this was to conduct a questionnaire survey supported by ADP analysis using the CODAP suite of programs.

4. The use of CODAP to analyse an information system and indeed a management organisation was to us an innovative extension of its capability. Its program OVLGRP (overlap and group) which clusters individuals together on the basis of similarity of task performance was considered to be potentially a very powerful tool for the job. Questionnaire design and data analysis were directed towards the use of this program.

QUESTIONNAIRE DESIGN

5. The major component of the 3-part questionnaire was a section which listed an inventory of data items potentially comprising the training information system. These were compiled from an examination of the present systems and from

experience within RNSETT where research and application into Computer Based Training had been established for 2 years. 105 elements were identified, classified into the following categories:

a. Trainee personal details
b. Class details
c. Course details
d. Module details
e. Examinations and tests
f. Training documentation
g. Future training requirement

6. Since the aims of the project were concerned principally with the introduction of a computer, the various elements of an ADP system were used to define activities applied to the information items. 5 functions were defined:

a. GATHERING - Collecting information from original source.

b. COMMUNICATING - Passing on information by any means.

c. STORING - Recording, filing, keeping information.

d. RETRIEVING - Obtaining information from files or records.

e. PROCESSING - Working on information to change its structure in any way.

7. To add to the refinement of the ADP system model it was decided to qualify involvement in any of the activities by a coarse measurement of the URGENCY attached to its performance and the EFFORT at present needed to achieve it. In addition for each information element a further response capability was added to enable respondents to indicate that the item was not readily available to them at present, but they would like it to be (NEED). So this part of the questionnaire developed into a matrix of rows of information elements X columns of activities, each activity split into URGENCY and EFFORT responses, with NEED in the final column. The 105 information items were thus expanded into 1155 response options. A copy of part of this section of the questionnaire is at Annex A.

8. It was of course appreciated that this was a complex format and to some extent experimental; this contributed to the aim of evaluating the system for future use. However, the task of the respondent was eased by only having to score 1 = yes, 2 = no in each applicable box and to make no response at all if he was not involved in the item/activity. Some weight was also given to the idea that the reason for the survey - the introduction of a management computer would provide an element of motivation to do a good job on the questionnaire!

9. The other two parts of the questionnaire were:

Part 1 - Respondent background information, and

Part 2 - Apportionment of Job Time

This latter section was included so that information usage could be compared with job profiles and comprised 51 training tasks against which the conventional

CODAP job description was computed from scores of 1-9 indicating relative time spent.

SAMPLE

10. In each of the 3 establishments just over 100 officers or senior ratings (312 in all) were identified as being significantly involved in training manageme Half of them were selected for the survey and completed questionnaires from 155 respondents were processed, distributed proportionately between the establishment

CODAP STRATEGY

11. The keystone of the CODAP strategy was the clustering program OVLGRP and the questionnaire and processing plan were built around it.

12. Two CODAP files were created from the original raw data file, the background variables being common to both. The processing plan is shown diagrammatically at Annex B. In file 1, responses to the URGENCY boxes were formatted as T (task time) factors; EFFORT and NEED responses as H's (history variables). We were aware that the range and meaning of responses to the URGENCY factors (1 = yes, 2 = no, 0 = not performed) could not be handled meaningfully by CODAP's job description (JOBDEC) computation. However, we had in effect a 2-point scale where a score of 1 or 2 indicated task performance. The job description calculation of members performing was therefore valid, and we were encouraged by the conclusions of PASS and ROBERTSON (NPRDC - 1980) that there would be littl loss of accuracy compared with percentage time results from a 5, 7 or 9 point scale. A measured statement of the relative performance of training information activities could therefore be produced easily from the JOBDEC program, by establishment or selected group, as illustrated at Annex C. This process established the parameters of the computer training information record.

13. Similarly, for the clustering program OVLGRP the TIME-OVERLAP formula could not be used, but the TASK-OVERLAP formula, which computes for each pair of individual job descriptions the average of the ratios of common tasks to total tasks individually performed, could be properly applied. The program was applied to the whole file and to sub-sets of the file for each establishment. In each case the resultant DIAGRAM was examined and nodal clusters were selected; for each of these groups programs were run to reveal group membership and "informatio usage profiles." Some of the profiles splayed remarkable bias towards one activity; groups could be identified as being GATHERERS, RETRIEVERS or STORERS. Annex D illustrates the point. The p ile of group 6 of the HMS FISGARD diagram shows that these 6 people have groupe ecause they are RETRIEVERS of information out of the first 75 activities in descending order of percent performing 65 are RETRIEVING.

14. Of course by the nature of the clustering formula we should not be too surprised at this result, but the clarity of job-typing was nevertheless most rewarding. We were able to identify the rank, organisational position and name of group members so that their similarity of information usage could be related to their function. This particular group comprised a Commander, 3 Lt Cdrs and 2 Lieutenants, most of whom could be classified immediately as being in positions in the management function where retrieval of information was a logical job featu

15. To come back to the strategy, membership of DIAGRAM groupings was used to select job descriptions from FILE 2's job time apportionment section. Using Group 6 from FISGARD again, we could see that the prime duty area is Administration and

the details of group task performance were readily obtainable from the JOBDEC printout.

16. Using both files the variable summary progam (VARSUM) was used to produce qualifying data on each information item on URGENCY, EFFORT and NEED so that deeper analysis of the system requirement could be carried out. The binary YES/NO response meant that examination of the modal scores could be carried out easily.

17. The 2-file methodology employed enabled us to produce dossiers for each of the selected groups. The dossiers comprised:

a. Group members by name.

b. Members' organisational positions.

c. A job profile histogram of the group and average overlap value of information usage.

d. Information involvement and activity factors.

e. Information activity - summary of significant features.

The dossier (for FISGARD Group 6 again) is reproduced at Annex E.

CONCLUSIONS

18. These are stated with reference to the aims of the project:

a. Identifying data for the computer record. The questionnaire and CODAP produced quickly and efficiently a listing of data items; priorities can be easily established and decisions taken on record length based on training data usage.

b. Developing the ADP System model. The CODAP clustering program and DIAGRAM provide an effective means of analysing the information system. There is no other method available to achieve comparable results within the timescale. Results would be improved if the whole target population completed the questionnaire.

c. Identifying duplication of information activity. Duplication can easily be identified through CODAP. However, it may not be significant except perhaps in the GATHERING function.

d. Identifying mis-matches between the information system and Organisational structure. The CODAP clustering technique and associated programs provide a unique method of identifying differences. Whether these are mis-matches can only be ascertained by further and deeper examination.

e. Finally - testing the questionnaire and CODAP in this role. The project was completed well within the 13 week time limit. The system has proved to be sound, efficient and rapid and it is considered unlikely that the job could have been done in any other way within the time scale.

RECOMMENDATION

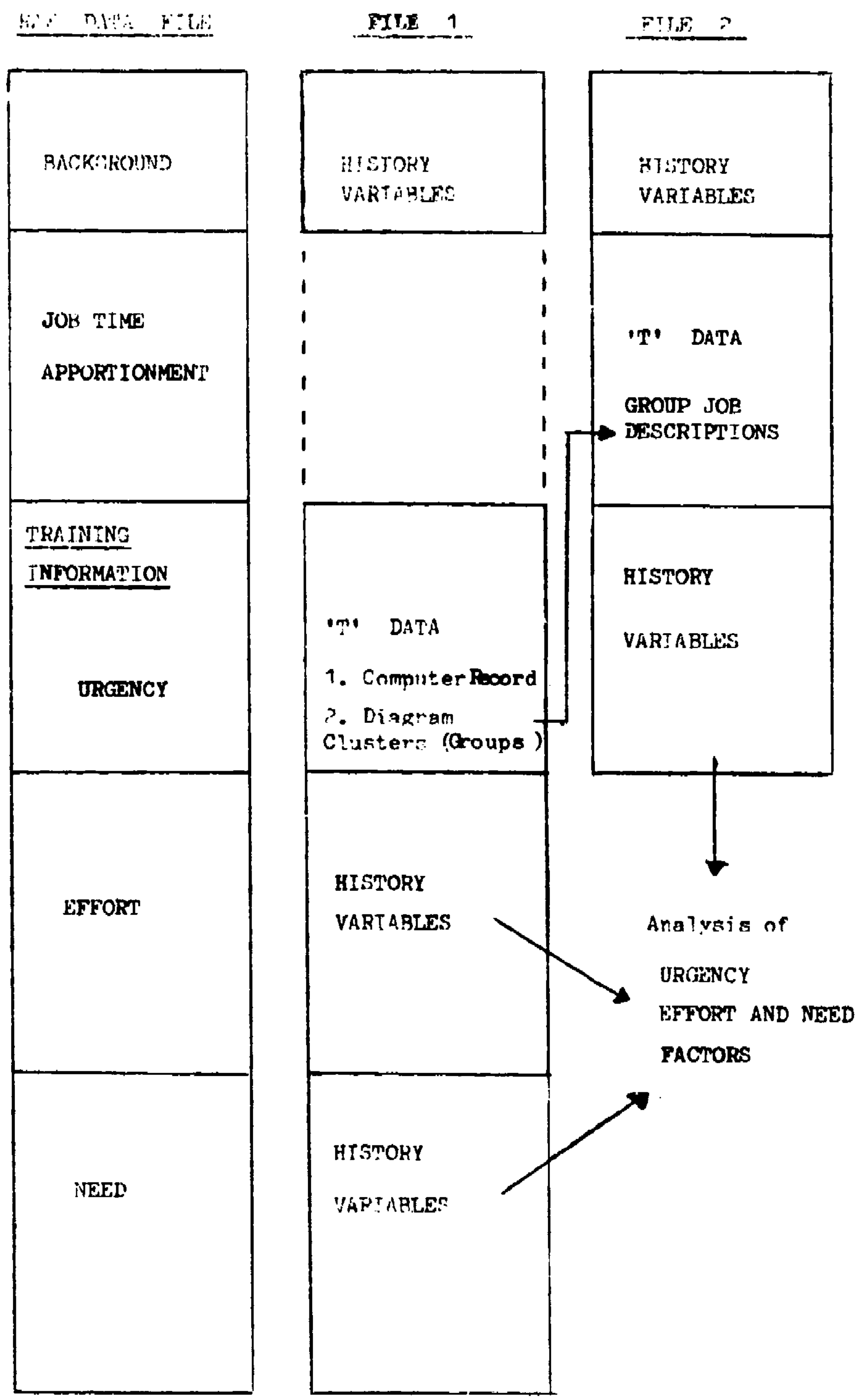
19. We have recommended to our management that the use of a CODAP based questionnaire for the recording and analysis of information systems should be placed high on the list of support facilities to be considered for any management survey in this field.

QUESTIONNAIRE EXTRACT - INFORMATION USAGE SECTION

Instructions Identify the information items you handle, then - choose the activity(ies) which best describe your involvement with the items and score 1 for YES and 2 for NO in the EFFORT and URGENCY boxes underneath each activity. If the item is not available to you and you think it should be, score 1 in the box underneath the NEED column. Otherwise leave boxes blank.

INFORMATION ITEMS	ACTIVITIES										
	GATHERING		COMMUNICATING		STORING		RETRIEVING		PROCESSING		NEED
	Urgency	Effort	Urgency	Effort	Urgency	Effort	Urgency	Effort	Urgency	Effort	
Course dates		18		20		22		24		26	27
Course Modules		29		31		33		35		37	38
Nominal list of trainees on course		40		42		44		46		48	49
Names of course manager, course instructors and administrators		51		53		55		57		59	60
Number of classes taking course		62		64		66		68		70	71
Course maximum loading 01 1 0 02	06	07		09		11		13		15	16
Number of courses per year		18		20		22		24		26	27
Individual trainee marks on course		29		31		33		35		37	38
Marks of all trainees on course (raw scores)		40		42		44		46		48	49
Historical record of course marks		51		53		55		57		59	60
Mean scores		62		64		66		68		70	71

CODAP PROCESSING PLAN



TRAINING MANAGEMENT COMPUTER RECORD - CODAP EXTRACT

HMS FISGARD

TASK JOB DESCRIPTION	CASES	TASKS	DUTIES	MEMBERS
	155	525	8	50

ORDERED BY PERCENT OF MEMBERS PERFORMING

TRAINEE NAMES AND FORENAMES	- STORING	74.0
TRAINEE NAMES AND FORENAMES	- RETRIEVING	68.0
DIVISION	- STORING	68.0
PRESENT CLASS	- STORING	66.0
DIVISION	- RETRIEVING	64.0
SERVICE NUMBER	- STORING	62.0
SERVICE NUMBER	- RETRIEVING	62.0
PRESENT CLASS	- RETRIEVING	62.0
BACK-CLASSING RECORD	- STORING	62.0

GROUP NUMBER-6 FOR ANALYSIS OF FISGARD HTK DIAGRM
RNSETT JOB NO 34/80

TASK JOB DESCRIPTION	CASES	TASKS	DUTIES	MEMBERS
	50	525	8	6

COUNT OF DUTIES OR TASKS LISTED
ORDERED BY
CUMULATIVE SUM OF AVERAGE PERCENT TIME SPENT BY ALL MEMBERS
AVERAGE PERCENT TIME SPENT BY ALL MEMBERS
AVERAGE PERCENT TIME SPENT BY MEMBERS PERFORMING
PERCENT OF MEMBERS PERFORMING

D-TSK	TASK TITLE		%
			****
A139	WARNINGS (TYPE, REASON, DATE, APPLIED/REMOVED)	-RETRIEVING	100.00
B 9	NOMINAL LIST	-RETRIEVING	100.00
C 49	INDIVIDUAL TRAINEE MARKS ON COURSE	-RETRIEVING	100.00
A 4	TRAINEE NAMES & FORENAMES	-RETRIEVING	83.33
A 69	DIVISION	-RETRIEVING	83.33
A 74	PRESENT COURSE	-RETRIEVING	83.33
A 79	PRESENT CLASS	-RETRIEVING	83.33
A134	EXAMINATION NUMBER	-RETRIEVING	83.33
A149	DAYS SICK	-RETRIEVING	83.33
A154	DAYS ABSENT	-RETRIEVING	83.33
B 24	CLASS DIVISIONAL OFFICER	-RETRIEVING	83.33
B 34	NUMBER IN CLASS	-RETRIEVING	83.33
A 34	SERVICE NUMBER	-RETRIEVING	83.33
A 59	TRAINEE NATIONALITY	-RETRIEVING	66.66
A 89	SPORTING INTERESTS	-RETRIEVING	66.66
A 94	HOBBIES RETRIEVING	-RETRIEVING	66.66
A 114	BACK CLASSING RECORD	-RETRIEVING	66.66

ANNEX E

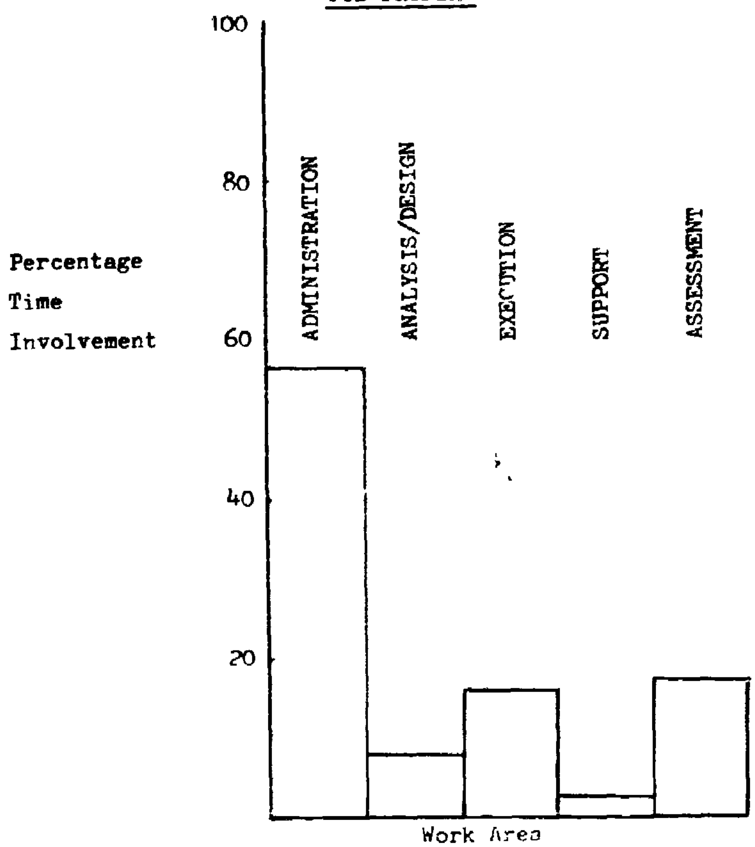
INFORMATION/JOB TIME DOSSIER-HMS FISGARD GROUP 6

FISGARD GROUP ANALYSIS

GROUP 6 (On Part 3 Training Information DIAGRAM)

Members: 6			Organisational Position
Commander	YOUNG	ADMIN	Technical Education: Craft
LT CDR	EVANS	ADMIN	Maths and Liberal Studies Officer
LT CDR	CHAPMAN	ADMIN	General Training Group Officer
LT CDR	LEWIS	ADMIN	Workshops Group Officer
LT	LAGES	EXECUTION	Maths Instructor

JOB PROFILE



DATA ACTIVITY FACTOR 172/525

PERCENTAGE OVERLAP: 51.9

DATA ELEMENT INVOLVEMENT	(85/105)	Potential
Trainee Personal Details	29	32
Class Details	8	9
Course Details	24	29
Module Details	10	10
Examinations and Tests	6	13
TEC	-	1
Six Part Documentation	5	6
Future Training Requirement	3	5

ACTIVITY NOTE

Out of the first 75 activities (in descending order of % performing), 65 are: "RETRIEVING"

Berkowitz, Melissa, US Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia. (Thurs. P.M.)

AREIS: A Computer-Based Educational Counseling System for the Army

Soldiers need information about career progression and educational options to maximize their professional development and their likelihood of remaining with the Army. Minimal information is provided by Army Continuing Education System counselors due to the volume of information needed to make career decisions and the limited time available to counsel each soldier.

This paper will discuss one solution to managing the vast quantities of career development information through the use of an individualized computer-based career counseling system. The US Army Research Institute, with contract support, has produced a prototype computer-based system called the Army Education Information System (AREIS). AREIS will assist the soldiers in defining work-related interests, skills and values to prepare them to identify their educational or vocational goals. The AREIS will also maintain a data bank for Education Center personnel so that career data can be compiled for planning and reporting purposes.

AREIS: A Computer-Based Educational Counseling System for the Army

Melissa S. Berkowitz, Ph.D.

Basic Skills Instructional Systems Technical Area
US Army Research Institute for the Behavioral and Social Sciences

AREIS: A Computer-Based Educational Counseling System for the Army[1]

INTRODUCTION

BACKGROUND

One objective of the US Army is to produce a combat-ready force through the development of personal skills and military proficiency. The Army Continuing Education System (ACES) supports this objective by providing educational opportunities to soldiers and enabling them to develop career goals that include military service and post-service education and training. Army Education Centers (AEC) have been established at every Army post having a minimum of 750 military members. These Centers provide programs which 1) satisfy the skill development and occupational needs of the Army, 2) increase soldier potential, 3) enhance job satisfaction, and 4) increase personal educational growth. Specific offerings for academic education include the Basic Skills Education Program (BSEP), Advanced Skills Education Program (ASEP), High School Completion Program, and Service members Opportunity Colleges Associate Degree (SOCAD) Program. In skill development the AEC provides language, military occupational speciality (MOS), and occupation-oriented courses. In the area of skill recognition the Army Apprenticeship Program and Defense Activity for Non-traditional Education Support (DANTES) Certificate Training are offered. The AEC also provided education services which include counseling, testing, and the support of a learning center.

The Adjutant General (TAG) supervises ACES and develops policy and guidance. Each installation/community commander conducts the ACES program through

1. This project was performed for the US Army Research Institute for the Behavioral and Social Sciences by the DISCOVER Foundation, Inc. under contract MDA-903-C-0279.

the Education Services Officer (ESO) so that educational and vocational opportunities described in the previous paragraph are made available to all service members. The ESO supervises the provision of these programs through contract instructors and the work of Education Center counselors. The counselors are required to provide each service member with program information and counseling during initial training, within 30 days of arrival at new duty stations, annually during the first enlistment, and 30 days prior to separation. The counseling emphasizes military professional development, educational opportunities, Veterans Assistance Program (VEAP) policies, and postservice educational benefits.

The primary means of delivering information about educational and vocational opportunities rests with the Education Center counselors. Two developments have hampered the provision of services by the Education Center staff: the increasing quantity and complexity of educational and vocational options, with a resultant explosion of resource information; and the reduction in the number of Education Center counselors. Education Centers, increasingly understaffed for the increased workload, are experiencing difficulty in adequately serving their constituency.

Hence it has become evident that other means of supplying standardized, up-to-date, easily accessible educational and vocational information are needed. One such means is a computer-based information system. Over the past two decades, a growing number of guidance professionals have become increasingly committed to the use of the computer to assist with the access and delivery of individualized educational and vocational information (Katz & Shatkin, 1980). The unique capabilities of the computer to store, search, retrieve, and update large masses of information; to relate educational and vocational data to information about the user; to simulate an interactive dialogue; and to serve many users simutaneously with tailored information have validated the worth of this technological aid to the counseling process.

The computer-based information system is intended to function in concert with, not instead of, the activities performed by guidance counselors. As the computer carries out information retrieving and dispensing functions and clerical duties, counselors would gain time to perform the professional duties for which they were trained and for which they are needed -- one-to-one interviewing, group guidance, and consultation.

This paper will discuss the design and on-going development of a computer-based educational and vocational information system which is one effort to overcome the increase of guidance information and decrease of counseling personnel in the military.

AREIS Army Education Information System

Needs Assessment. The US Army Research Institute for the Behavioral and Social Sciences (ARI) initiated a research effort to conceptualize and develop a prototype computer-based system which would provide information on military and civilian education programs related to the Army career progression. This effort, performed by the DISCOVER Foundation, Inc., under contract MDA 903-79-C-0279, provided a design for the Army Education Information System (AREIS) based on the results of a needs assessment survey administered to Education Services Officers (ESOs) and Education Center counselors at posts worldwide (Harris-Bowlsbey & Raybush, in press).

The needs assessment instruments were designed to collect data concerning 1) demographic information about the Education Center, 2) the variety and frequency of information requested by soldiers at the Education Center, and 3) ESO and counselor attitudes about using computers. The instruments were distributed to all major commands. The return rate for the ESOs was 72% with 131 of 182 questionnaires mailed back. The return rate for counselors was 64% with 313 of 494 counselors responding.

The following summarizes the demographic data supplied by the ESOs and counselors. Of the 144 posts responding the permanent population ranged from 50 to 48,000. The number of counselors per post ranged from 2 to 12. Data indicated that each counselor annually serves between 1,000 and 2,000 soldiers. The average workload is 1,600 soldiers per counselor.

Counselors indicated that half of their time is spent on one-to-one counseling of soldiers with the remainder distributed over administrative duties, orientation/outreach programs, clerical duties, liaison efforts, research and development, and other miscellaneous tasks. Counselors provide an average of two interviews per soldier per year. This represents approximately 64% of their workload. Counselors and ESOs ranked tuition assistance, college course offerings on or near post, and information about tests (DANTES, SAT, CLEP) as ACES program information they were most frequently asked about. In declining order of frequency they were asked about orientation to the Education Center services, associate degree programs, college credit for military experience, and BSEP. In ranking information requested about career planning, counselors and ESOs indicated the following in descending order of frequency: developing a personal career plan in and beyond the military, assessing interests, and making the transition from a military to a civilian job.

ESOs and counselors also responded to a series of questions to determine their attitudes about the usefulness of computerization of ACES information now available in print form. They indicated that computerization of information about new and existing ACES programs, Department of the Army regulations, master schedule of courses, and MOS and civilian occupations would be considerably useful. Counselors and ESOs agreed that a computerized system would provide soldiers with consistent information and would most likely be used frequently by soldiers. They also agreed that this type of system would be welcomed by counselors because it would enable them to counsel more soldiers by reducing their administrative workload. Counselors also indicated a need for training on the use of a computerized system. In general, counselors and ESOs were positive about the usefulness of this type of system as a tool to support Education Center operations.

AREIS Specifications. The results of the needs assessment provided DISCOVER Foundation, Inc. with the data necessary to formulate a conceptualization of a computer-based guidance system designed specifically for Education Center use. The AREIS is composed of four interactive subsystems. Subsystem I is the ORIENTATION which is the entry point for the soldier. The objectives of this subsystem are to 1) familiarize the user with the computer terminal and printer, 2) provide instruction about the content of the AREIS, 3) explain the Education Center services and 4) provide an overview of all ACES programs. The second subsystem is SELF-INFORMATION which has been designed to help soldiers generate information about themselves to formulate short or long range goals for their active duty and beyond. Subsystem II helps soldiers define their work-related interests, aptitudes, skills and values. Subsystem III GOALS and PLANNINC helps soldiers identify goals related to career and education and provides details of ACES programs which can help them achieve their goals. This subsystem provides information on the following goals:

1. to improve basic skills
2. to develop new interests for self-improvement or use of leisure time
3. to get some job skills
4. to complete the next step in education
5. to plan a military career

6. to improve MOS proficiency

7. to select a secondary MOS

8. to get promoted

9. to make a good decision about re-enlistment

10. to make a vocational choice

11. to complete an educational degree after leaving the military

12. to make the Army a career

Subsystem IV COUNSELOR-ADMINISTRATOR has been designed to reduce the clerical workload of counselors and provide them with information to be used during counseling interviews. The subsystem contains data files which include descriptions of MOSs, civilian occupations, and educational opportunities. These files may be accessed directly by the soldier through interactive dialogue or by the counselor. A second part of this subsystem, accessible only by Education Center personnel contains the Soldier Educational Development Record (DA Form 669) for each soldier, a master schedule of courses offered on or near the post, and all course rosters.

AREIS Field Tryout. A field tryout of portions of the four AREIS subsystems was conducted at the Ft. Sill, Oklahoma Education Center in April 1980. The following segments, which represent approximately one-third of the total system as specified above, were tested:

1. Subsystem I ORIENTATION: an overview of AREIS, Education Center services and ACES programs.

2. Subsystem II SELF-INFORMATION: on-line administration of the UNIACT Inventory (c 1978, American College Testing Program). This is a sixty item interest inventory which provides the respondent with a family of occupations to examine.

3. Subsystem III GOALS and PLANNING: the goal entitled "To complete the

next step in Education," designed to provide information about educational offerings on or near a specific Army post.

4. Subsystem IV COUNSELOR-ADMINISTRATOR: a demonstration of administrative documents which may be maintained by computer such as DA Form 669, master schedule of courses, and summary report data.

The preceeding segments of the AREIS subsystems were programmed in PLANIT (Programming Language for Interactive Teaching) on the Army's UNIVAC 1108 Computer at the Edgewood Arsenal, Maryland and delivered to Ft. Sill in a time sharing mode.

Twelve counselors and sixty-four soldiers participated in the field tryout. The soldiers were volunteers who had come into the Education Center for information. On-line surveys were given to the soldiers prior to using the AREIS and after each subsystem to determine their attitudes on the usefulness, clarity, and interest level. The computer and the AREIS content were perceived useful by the soldiers for educational and vocational planning. Counselors indicated that the information provided by the AREIS subsystems was useful and accurate. They responded favorably to the delivery of educational information to soldiers by computer.

Future Direction for AREIS. Recently a contract was awarded to the DISCOVER Foundation, Inc. to complete the AREIS subsystem development and conduct a field trial of the system at three Army sites. A cost/benefit analysis of alternate delivery systems for the AREIS, conducted under the previous contract, has guided the selection of a micro-computer for the AREIS hardware. The hardware systems compared for the cost/benefit analysis were the maxi-computer, distributed network of mini-computers, and stand-alone micro-computer. The micro-computer was recommended because it has the greatest cost feasibility, requires a minimum of technical and clerical support, is easily operated by non-technical personnel, and can be readily installed overseas. The AREIS hardware will consist of a microcomputer, color monitor, and printer. Some limitations of the PLANIT authoring language were identified during the field tryout. These include the inability to search data files, clear the screen completely, and remain in contact with the computer after a five minute delay between users. Another authoring language will be selected for the software development which does not have the above limitations.

AREIS support documentation will be prepared to provide a User's Manual and a program for in-service training of ESO's, counselors, and clerical personnel. AREIS subsystem software will be completed according to the conceptualization discussed previously. An in-depth field trial of the total system will be conducted at three Education Centers: Ft. Meade, Maryland; Ft. Gordon, Georgia; and Heidelburg, Germany. The field trial will provide data on the use patterns of all subsystems, user reactions, influence of system use on soldiers, and impact of the system on ESOs, counselors, and clerical staff. This effort will be completed during the second quarter of FY 83.

CONCLUSION

ARI is guiding the development and field trial of the AREIS as one solution to the problem of the surge of educational and vocational information being provided by a decreasing number of Education Center counselors. A preliminary field tryout has indicated that the application of computer technology to Education Center operations is most welcome by ESOs, counselors, and soldiers. The potential payoff for the AREIS may be observed in increased soldier potential, job satisfaction, and personal educational growth while supporting the occupational needs of the Army in the defense of the nation.

REFERENCES

Harris-Bowlsbey, J. & Raybush, C. Army continuing education system needs assessment (Technical Report). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, in press.

Katz, M. R. & Shatkin, L. Computer assisted guidance: concepts and practices (ETS Research Report 80-1). Princeton, NJ: Educational Testing Service, March 1980.

AD P001292

Item Response Time as a Basis for Ability and Difficulty Measures

Werner Birke
Federal Armed Forces Office
(Germany)

Summary

For the purpose of adaptive intelligence measurement (CAT) a model is proposed, which uses response-time data for estimating person parameters and item parameters on one scale. It is assumed that there is a close connection between two relations: ability and difficulty on the one hand, response-time and critical response-time (CRT) on the other hand. The CRT of an item is defined as the mean RT of subjects whose ability equals the difficulty of this item. In such test situations the probability of a correct answer equals .5. Relevant subject-item-combinations (i.e. critical response times) can be identified by rank-ordering the lines and columns of a solution-matrix according to sums correct. The border (diagonal) between the correct and incorrect triangle of the matrix is the guide-line of the model.

Explanation of the model

Adaptive testing makes no use of the sum of correct answers, because most of the subjects have the same sum correct, derived from different item samples. A single correct (or incorrect) answer gives us only a very rough bit of information. In fact it tells us nearly nothing. It needs very complicated scaling methods (as provided for instance by item-response-theory) to derive estimates of person-parameters from binary data. I propose to make use of an additional source of data: of item response-time.

Dependent variables like intelligence scores may be determined by a host of different independant variables. Most of these we don't want to be mixed up with variables which we try to measure. Because in practice we can not identify and substract the so-called error variance from the empirical scores and get true scores, we might as well forget about the error variance in the phase of test construction and test application and take account of it in the phase of test interpretation.

For the measurement itself I assume that an item-response x of subject s to item i is merely a function of the relation between the ability A of subject s and the difficulty D of item i.

$$x_{si} \longrightarrow f\ (A_s\ /\ D_i)$$

From a correct answer ($x_{si} = 1$) I conclude that ability A_s is greater than difficulty D_i. And from an incorrect answer ($x_{si} = 0$) I conclude that ability A_s is smaller than difficulty D_i.

$$(x_{si} = 1) \longrightarrow A_s > D_i$$
$$(x_{si} = 0) \longrightarrow A_s < D_i$$

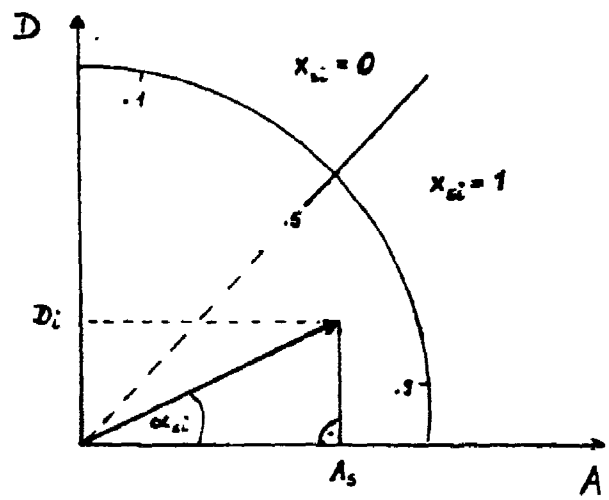
With a given subject-item-combination, the probability of a correct response depends on the relation between the respective ability and difficulty.

$$p_{si} \longrightarrow f\,(A_s \,/\, D_i)$$

The probability p_{si} increases with increasing ability and/or decreasing difficulty.

Subject and item are independent components of the test situation. Therefore ability and difficulty may be described as ordinate and abscissa, or as short sides of a rectangular triangle.

Fig. 1



In Figure 1 the relation between A_s and D_i is the cotangent of angle α_{si}

$$\text{cotg}\ \alpha_{si} = A_s \,/\, D_i$$

$$\alpha_{si} = 90\ (1 - p_{si})$$

The smaller the angle the greater the probability of a correct answer.

The basic assumption of my model says, that there is a correlation between item-response-time and angel α_{si}.
In other words:

> The response time t_{si} needed by subject s to solve item i is a function of the relation between the ability A_s and the difficulty D_i

$$t_{si} \longrightarrow A_s / D_i$$

$$t_{si} \longrightarrow \operatorname{cotg} \alpha_{si}$$

$$t_{si} \longrightarrow \operatorname{cotg} 90\,(1-p_{si})$$

The most simple relation between item-response-time and probability of a correct answer would be a linear function. But I don't expect things to be as simple as that in reality (see Figure 2).

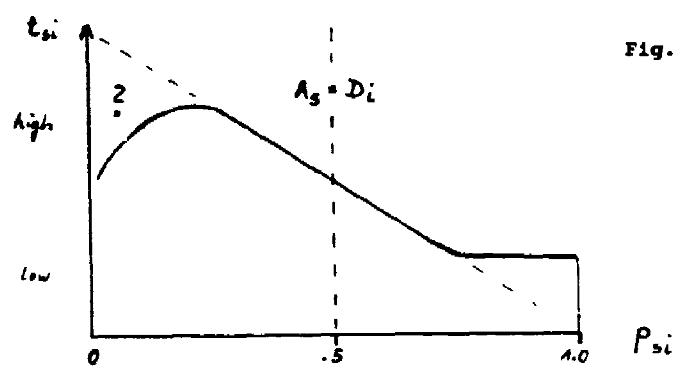


Fig. 2

To every item there is a minimum response-time which may perhaps be reached at a probability level of .8 or higher. If on the other hand the probability of a correct answer is very low, i.e. if the item is much too difficult for the subject, response-time will in many cases be relatively low, because the subject realizes that he has no chance and gives up or guesses. So only in the middle region, where the probability of a correct answer is about .5, I expect t_{si} to be a linear function of p_{si}: the higher the probability the shorter the response-time. Suppose now that we have a second subject c but the same item. At least with medium probability of a correct answer (CAT-region) I expect the following equations to hold true:

$$\frac{A_s/D_i}{A_c/D_i} = \frac{p_{si}}{p_{ci}} = \frac{\operatorname{cotg}\alpha_{si}}{\operatorname{cotg}\alpha_{ci}} = \frac{\alpha_{si}}{\alpha_{ci}} = \frac{t_{si}}{t_{ci}}$$

Let subject c be a special subject; i.e. a subject whose ability equals the difficulty of the item.

$$A_c = D_i$$

In this case the angle α_{ci} (Fig. 1) will be 45°.

$$\frac{\alpha_{si}}{45^{\circ}} = \frac{t_{si}}{t_{ci}}$$

$$t_{ci} = t_{.i}$$

Appointment:

The critical response-time $t_{.i}$ of item i, is the mean response-time of those subjects whose ability equals the difficulty D_i. In such cases the probability of a correct answer to item i is equal to .5.

$$\frac{\alpha_{si}}{45} = \frac{t_{si}}{t_{.i}} \qquad \alpha_{si} = 45\,\frac{t_{si}}{t_{.i}}$$

The relation between the individual response-time of subject s answering item i and the critical response-time of item i determines the size of angle α_{si}. The cotangent of this angle is the relation between the ability of subject s and the difficulty of item i.

$$\boxed{\frac{A_s}{D_i} = \text{cotg } 45 \frac{t_{si}}{t_{.i}}}$$

(Formula 1)

A = ability (ordinate)
s = subject
D = difficulty (abscissa)
i = item
t_{si} = response-time
$t_{.i}$ = critical RT

For practical use this equation can be transformed into the following versions:

$$A_s = D_i \ . \ \text{cotg } 45 \frac{t_{si}}{t_{.i}} \qquad \text{(Formula 2)}$$

$$D_i = A_s \ . \ \text{tg } 45 \frac{t_{si}}{t_{.i}} \qquad \text{(Formula 3)}$$

Critical response-time (CRT)

The critical RT of an item was defined above as "the mean RT of those subjects whose ability equals the difficulty of the item". How can we determine CRT, if we do know neither ability nor difficulty?

In test situations where ability and difficulty are equal, the probability of a correct answer is equal to .5. This we can estimate. The probability p_{si}, that subject s will solve item i

correctly, has two aspects. There is

- an active probability $p_{s.}$ of subject s to solve correctly, and
- a passive probability $p_{.i}$ of item i to be solved correctly.

In a solution matrix these two aspects correspond to the relative frequencies of answers correct in lines and columns of the matrix.

$$p_{si} = \frac{p_{s.} + p_{.i}}{2} \qquad p_{s.} = (\sum_i x_{si}) : n_i$$

$$p_{.i} = (\sum_s x_{si}) : n_s$$

The following check-list outlines the procedure for estimating the critical response-times for the items:

- item-sample and subject-sample should at best be evently distributed in respect to ability and difficulty
- give all items one by one in random order to all subjects on a screen
- measure the time from item-presentation to key-pressing
- make out a solution-matrix and a response-time-matrix
- rearrange the lines and columns of the solution-matrix so that they both form a rank-order according to sum correct (see simplified example next page)
- rearrange the t_{si}-matrix in the same way as the x_{si}-matrix
- make out the p_{si}-matrix ($p_{si} = (p_{s.} + p_{.i}) : 2$)
- compare p_{si}-matrix with t_{si}-matrix and find in every item-column (if necessary by interpolation) the $t_{.i}$-estimate which corresponds to $p_{si} = .5$; if there are two or more relevant estimates for $t_{.i}$, compute the mean.

solution-matrix

s \ i	1	2	3	4	5	6	7	f_s
1	+	-	+	-	+	+	-	4
2	+	-	+	-	+	-	-	3
3	+	+	+	+	+	+	-	6
4	-	-	+	-	-	-	-	1
5	+	-	+	+	+	+	-	5
6	+	+	+	+	+	+	+	7
7	-	-	+	-	+	-	-	2
f_i	5	2	7	3	6	4	1	

(t_{si})

response-time-matrix

s \ i	1	2	3	4	5	6	7
1	3.3	8.9	4.3	5.4	2.6	5.5	8.4
2	4.0	9.6	4.9	6.0	3.0	6.3	10.4
3	2.6	7.5	4.1	4.2	2.6	4.4	7.0
4	5.8	10.6	6.1	7.5	4.1	7.9	9.6
5	2.9	8.2	3.9	4.8	2.5	4.8	7.6
6	2.7	6.8	3.8	4.1	2.4	4.3	6.4
7	4.9	10.8	5.4	6.7	3.5	7.1	10.5

rearranged
solution-matrix

s \ i	3	5	1	6	4	2	7	f_s	$p_{s.}$
6	+	+	+	+	+	+	+	7	1.0
3	+	+	+	+	+	+	-	6	.86
5	+	+	+	+	+	-	-	5	.71
1	+	+	+	+	-	-	-	4	.57
2	+	+	+	-	-	-	-	3	.43
7	+	+	-	-	-	-	-	2	.29
4	+	-	-	-	-	-	-	1	.14
f_i	7	6	5	4	3	2	1		
$p_{.i}$	1.0	.86	.71	.57	.43	.29	.14		

p_{si}-matrix

s \ i	3	5	1	6	4	2	7	$p_{s.}$
6	1.0	.93	.86	.79	.71	.64	.57	1.0
3	.93	.86	.79	.71	.64	.57	.50	.86
5	.86	.79	.71	.64	.57	.50	.43	.71
1	.79	.71	.64	.57	.50	.43	.36	.57
2	.71	.64	.57	.50	.43	.36	.29	.43
7	.64	.57	.50	.43	.36	.29	.21	.29
4	.57	.50	.43	.36	.29	.21	.14	.14
$p_{.i}$	1.0	.86	.71	.57	.43	.29	.14	

rearranged RT-matrix

s \ i	3	5	1	6	4	2	7
6	3.8	2.4	2.7	4.3	4.1	6.8	6.4
3	4.1	2.6	2.6	4.4	4.2	7.4	7.0
5	3.9	2.5	2.9	4.8	4.8	8.2	7.6
1	4.3	2.6	3.3	5.5	5.4	8.9	8.4
2	4.9	3.0	4.0	6.3	6.0	9.6	10.4
7	5.4	3.5	4.9	7.1	6.7	10.8	10.5
4	6.1	4.1	5.8	7.9	7.5	10.6	9.6
$t_{.i}$	7	4.1	4.9	6.3	5.4	8.2	7.0

Item calibration

Once we have computed the critical response-time-values of the items, we can make out the α-matrix ($\alpha_{si} = 45\ (t_{si} : t_{.i})$), on the basis of which ability and difficulty can be estimated by use of formulas 2 and 3. (See the continuation of our example, next page). But at first we must decide where the parameter-scales are to be anchored.
In our example we might arbitrarily state that subject No. 1 has an ability-score of $A_1 = 100$. Mathematically it would be possible to compute nearly all difficulty-scores by use of response-times or alpha-values in the matrix-line of subject No. 1. (Only item 3 can not be calibrated, because we have no $t_{.3}$-value). But it is not advisable to do it this way.

For the purpose of item-calibration we should make use only of those test-situations, where the respective item was given to a subject whose ability is about as high as the difficulty of the item and who has solved the item correctly. As mentioned above, I expect response-time data to be relatively invalid indicators for the ability/difficulty-relations, if there is a great difference between A_s and D_i. So only the t_{si}-values in the diagonal region of the matrix should be used. That means that we should start with the anchor subject (or anchor-item) and go along the diagonal into both directions, estimating alternatively A-values and D-values.

Application

The model was worked out for computerized adaptive testing (CAT). It makes it possible to estimate a subjects ability on single items. A calibrated item represents a point on the difficulty scale. The ability of a subject is assumed to be greater or smaller than this value, depending on whether the answer is correct or incorrect. This information can be specified by taking response-time into account and estimating ability with formula 2. So the item does not only define an ability limit, but measures a certain ability-range adjacent to its difficulty-value. Therefore item-pools for the practice of CAT can be smaller than they are required for other methods used today. I hope that this statement will be confirmed by empirical investigations.

α-matrix

s \ i	3	5	1	6	4	2	7
6		26.3	24.8	30.7	34.2	37.2	41.1
3		28.5	23.9	31.4	35.0	40.6	45.0
5	2	27.4	26.6	34.3	40.0	45.0	48.9
1		28.5	30.3	39.3	45.0	48.8	54.0
2	•	32.9	36.7	45.0	50.0	52.7	66.9
7		38.4	45.0	50.7	55.8	59.3	67.5
4		45.0	53.3	56.4	62.5	58.2	61.7

$A_s = D_i \cdot \mathrm{cotg}\,\alpha_{si}$ $\qquad$ $D_i = A_s \cdot \mathrm{tg}\,\alpha_{si}$

$A_1 = 100$

$D_6 = A_1 \cdot \mathrm{tg}\,\alpha_{16} = 100 \cdot \mathrm{tg}\ 39.3 = 82$

$D_4 = A_1 \cdot \mathrm{tg}\,\alpha_{14} = 100 \cdot \mathrm{tg}\ 45.0 = 100$

$A_5 = D_4 \cdot \mathrm{cotg}\,\alpha_{54} = 100 \cdot \mathrm{cotg}\ 40.0 = 119$

$D_2 = A_5 \cdot \mathrm{tg}\,\alpha_{25} = 119 \cdot \mathrm{tg}\ 45.0 = 119$

$A_3 = D_2 \cdot \mathrm{cotg}\,\alpha_{32} = 119 \cdot \mathrm{cotg}\ 40.6 = 139$

$D_7 = A_3 \cdot \mathrm{tg}\,\alpha_{37} = 139 \cdot \mathrm{tg}\ 45.0 = 139$

$A_2 = D_6 \cdot \mathrm{cotg}\ _{26} = 82 \cdot \mathrm{cotg}\ 45 = 82$

$D_1 = A_2 \cdot \mathrm{tg}\ _{12} = 82 \cdot \mathrm{tg}\ 36.7 = 61$

$A_7 = D_1 \cdot \mathrm{cotg}\ _{71} = 61 \cdot \mathrm{cotg}\ 45 = 61$

$D_5 = A_7 \cdot \mathrm{tg}\ _{75} = 61 \cdot \mathrm{tg}\ 38.4 = 48$

Item	1	2	3	4	5	6	7	
Diffic.	61	119	?	100	48	82	139	

The Maintenance Performance System

Dr. Douglas J. Bobko
Research Psychologist
Army Research Institute
Alexandria, Virginia

The Maintenance Performance System (MPS) is a computer-based management system designed to improve the conduct and quality of maintenance training at the direct support level; it has been in operation at a FORSCOM maintenance battalion for nearly one year. An evaluation of the system suggests that MPS monitors daily maintenance activities with relative ease and accuracy. Moreover, MPS-generated skill and training information is used by shop personnel to diagnose skill deficiencies and to guide technical training of individual repairmen.

Background

The proper and continuous maintenance of Army vehicles and other equipment is crucial to achieving combat readiness. However, a number of reports have indicated that, at both the organizational and direct support level, equipment maintenance procedures are less than optimal. Typical of recurring problems in maintenance operations are these: faulty diagnosis, failure to recognize equipment failures, unnecessary part replacement, and failure to conduct on-the-job training.

In an effort to help correct these maintenance deficiencies, and with the assumption that effective maintenance training helps ensure efficient maintenance operations, the Army Research Institute (ARI) developed the Maintenance Performance System (MPS).

The Maintenance Performance System

The primary goal of MPS is to improve the conduct and quality of on-the-job training of repairmen in their performance of technical tasks at direct support battalions (a more extensive system for use at the organizational level is now being designed). To accomplish this, MPS provides unit management with current information about the experience, training and skills of maintenance personnel, and then guides the manager toward selection of available training methods and materials to meet skill and performance needs.

Integration Into Normal Operations

The structure of MPS and its integration into normal operations are shown in Figure 1. Daily changes in both equipment status and repairmen experience are recorded and fed into a computer file; in turn, information about shop performance on an aggregate and individual level is prepared in report form. These reports from MPS are given to battalion and company level managers who use them to diagnose maintenance performance deficiencies. Managers can then use specific system-derived guidance to match training resources with training needs and to develop an on-the-job training program.

It is important to note that MPS differs from the current Army Maintenance Control System (MCS). MCS is primarily designed to provide information about the status of equipment as it passes through the shop and, unlike MPS, MCS provides no information about individual and unit skills, experience, or training.

Description of System

MPS has been in use for nearly one year by two forward support companies of a divisional FORSCOM maintenance battalion. It currently includes ten Military Occupational Specialties (MOS) and almost all of the equipment serviced by the

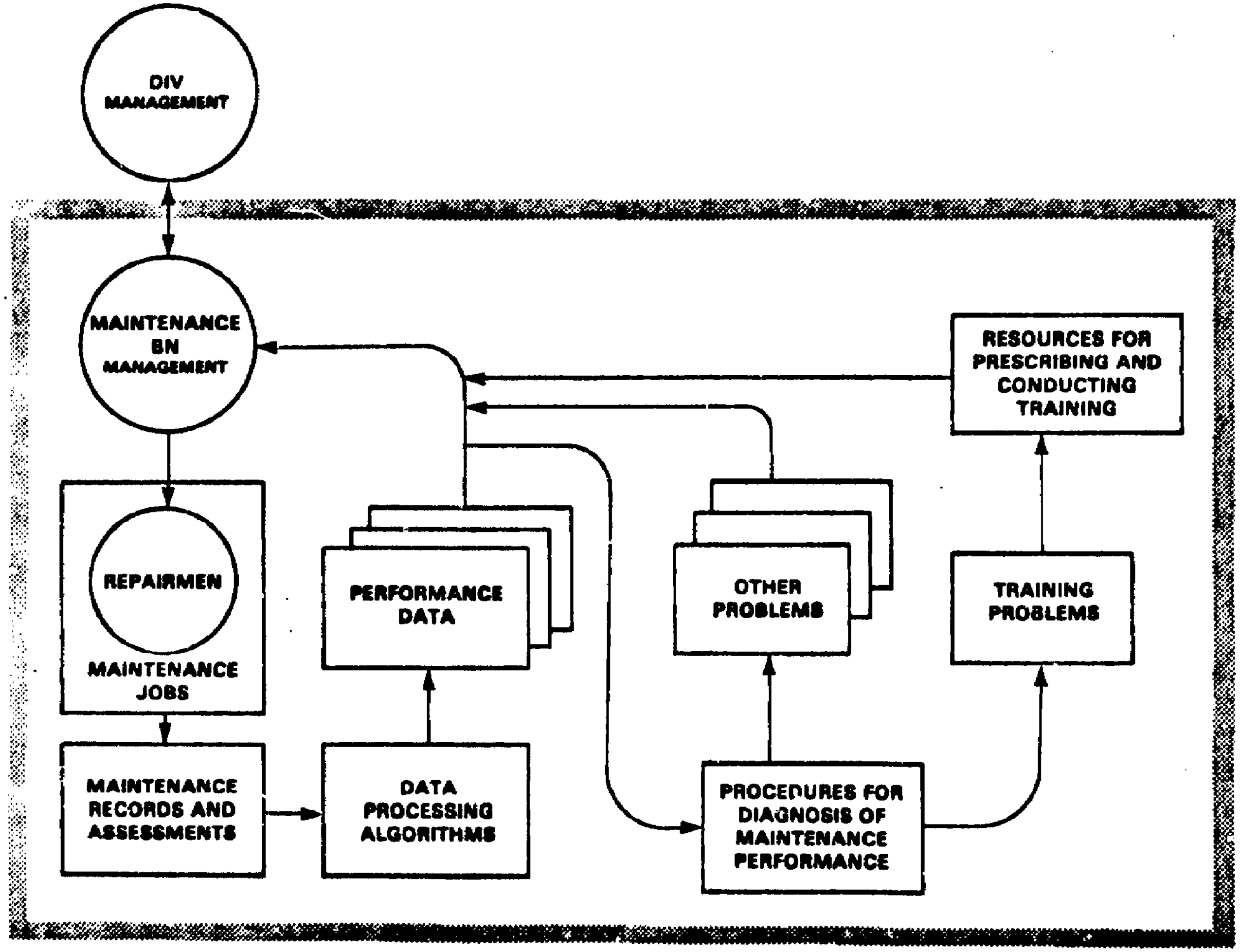


Figure 1. Integration of the Maintenance Performance System into maintenance operations.

companies. The MOSs include the high density MOSs of 63H (track vehicle repairmen), 63W (wheel vehicle repairmen), and 45K (tank turret repairmen).

Technical MOS supervisors feed information into the system through a set of five simple input forms. Based on observations to date, supervisors spend about ten minutes each week completing them. The information required on each form is described in Table 1.

TABLE 1. DATA COLLECTION FORMS

Input Form	Title	Use
1	Job Order Status	One for each job; records time of each change of job status
2	Job Performance	Records amount of time each repairman works on each job
3	Daily Man-Hour Availability	For each repairman, records available hours, direct hours and overtime
4	Training/Performance Demonstration	Records accomplished training or testing for each repairman
5	Task Experience	Repairman entering system report prior experience on each task

Day to day operation of MPS is accomplished by a junior enlisted clerk in grade E3 or E4. The clerk is responsible for collecting the input forms, key encoding the data into the computer (presently an IBM 5120), generating computer-printer reports, and distributing the reports. The clerk works full time at these duties.

In all, there are nine different MPS reports and two tables (see Table 2). One table lists the complete roster of each company and is distributed to the company office. It is used to keep track of the number and type of personnel in the unit on a bi-weekly basis. The other table contains a list of comments about local conditions that may affect the interpretation of reports. For example, this table would record when training holidays occur, because training holidays would affect the time and direct labor available for maintenance.

There are 5 different reports which contain data that reflect the operational status of the unit and these are delivered to relevant personnel on a bi-weekly schedule. Maintenance managers use information about man-hour use, time-to-repair, and time in each job status (i.e., awaiting parts, in-shop, etc.) to locate deficiencies in maintenance performance and other problems. Maintenance managers are guided in the diagnosis of problems by an "Interpretation Guide" that is specifically keyed to MPS.

Four other reports, which are produced every six weeks, contain data that are particularly relevant to the analysis of training needs and the guidance of training activities. For example, the Individual Skill History report contains a record of each individual's experience on each MOS technical task. Experience on a task includes the job exposure that occurs when a soldier is given a task to perform, more formal training that is provided to train or refresh skills, and demonstrated proficiency on a performance-based test. The Individual Skill History report is essentially an automated procedure for recording the tasks that each soldier performs and is much like a job book. It also serves as the basis for the Training Requirements Summary report which actually guides training activity. This report lists those individuals who have insufficient experience on each task and ranks the need for training according to the degree of experience and the criticality of the task. Training managers can use this report to identify those individuals who most need training on each task.

TABLE 2. MPS REPORT INFORMATION

Table 1	Roster
Table 2	Interpretation Comments
Report 1	Man-hour Availability and Use
Report 2	Average Direct Man-hours Per Job
Report 3	Average Direct Man-hours Per Job by Equipment and Task
Report 4	Average Job Throughput Time in Days
Report 5	Average Days Spent in Each Job Status
Report 6	Skill and Growth Indices
Report 7	Skill Development Summary
Report 8	Individual Skill History
Report 9	Training Requirements Summary

Evaluation of MPS

MPS has been in operation for several months withour the assistance of or supervision by ARI. Several visits and observations during this time have revealed that MPS functions smoothly on a day-to-day basis, e.g., forms are routinely attached to job orders and filled in, data is collected and entered into the computer regularly, and reports are printed and distributed appropriately. A more detailed analysis of MPS was conducted, however, to evaluate the accuracy and validity of MPS information, and to determine the extent to which MPS training reports affect the conduct and content of skill training in the maintenance shops.

The first objective of this evaluation was to determine the accuracy of information being entered onto the more critical MPS forms. This was done in light of the fact that MPS report interpretation and consequent command action is predicated on confidence in the accuracy of MPS statistics; data input quality is a fundamental problem for any such system.

A second objective of the evalution was to assess the validity of MPS-generated individual skill indices. The skill index, which is assigned to each repairman, can range from 0 to 100 and it reflects the degree to which a repairman is proficient on those tasks covered by his MOS; the skill index increases each time a repairman works on a job, receives training, or passes a performance-based test on a job. (The skill index is essentially a single number describing the content of the Individual Skill History report.

Skill indices are an integral part of MPS for a variety of reasons. For battalion and company commanders, they point to current strengths and weaknesses in overall maintenance capability, and provide a measure of the growth in maintenance proficiency. For shop supervisors, the skill index is an easy reference to those repairmen who require additional training. To the individual repairman who receives a skill growth report every six weeks, the skill index can serve as a source of reward and motivation.

The third objective of this evaluation was to determine the extent to which MPS reports of job experience are used for job assignments and for setting training priorities.

Accuracy of MPS data

Most of the critical information monitored by MPS is collected on MPS Forms 1, 2, and 3. Hence, a sample of these forms in each company was examined and checked for accuracy.

MPS Form 1 (Job Order Status). For two companies and four MOSs, a total of 67 forms were examined. The information contained on these forms was compared with similar information contained in the job folder which normally follows a job through the shop; shop officers were also questioned about data entries on this form. It was found that the information contained on MPS Form 1 is generally consistent with that of DA Form 2407 and with verbal reports from shop officers. The most frequent problem was an incorrectly entered or omitted job code; this occurred on six of the forms examined.

MPS Form 2 (Job/Task Performance). A total of 18 Form 2s were chosen across the two companies and for four MOSs; at the time of their selection, these forms represented shop work completed within the last 48 hours. The repairman (and their supervisors) listed on these forms were questioned about the jobs they had recently performed, and the length of time each job required. In each case, the names listed on Form 2 were consistent with reports from repairmen and supervisors, i.e., those repairmen who performed the job received credit for it.

The accuracy of job completion time data was more difficult to determine, however. Since there are occasional interruptions in work, and because start and stop times are not recorded with a stopwatch, job completion times are generally rough estimates. Based on reports from repairmen and supervisors, the total job completion time is typically rounded up to the nearest hour or half hour, so that the magnitude of error increases as the job completion time decreases. For this sample of forms, however, there was no instance in which repairmen's reports of job completion time differed from MPS data entries by more than 60 minutes.

MPS Form 3 (Daily Man-Hour Availability). Fifty-five forms were reviewed, which represent both companies and four MOSs. From interviews with shop supervisors, it is clear that available man-hour entries are accurate but that the direct man-hour entries are predominantly inaccurate. For example, it is a common practice with Form 3 to enter 7 or 8 direct man-hours for repairmen who worked only 4 or 5 hours. That direct man-hour entries are inflated was confirmed by a cross-check for Form 3 with Form 2 (which records job performance time for each repairman).

Validity of MPS Skill Index

To assess the validity of the MPS skill index, the indices were compared with independently obtained job performance ratings of each repairman. In short, each repairman tracked by MPS was rated by both his supervisors and his peers. Ratings were made with the use of an anchored scale, where a 0 performance rating indicated a repairman who required constant and total supervision on a job, 50 indicated that some supervision was required, and a performance rating of 100 meant that the repairman could perform a task successfully without any supervision. Three ratings were obtained for each repairman: the first rating was to indicate a repairman's general performance in his MOS, and the remaining two ratings were to indicate his performance on two different but specific tasks (which differed for each MOS). In all, usable ratings were obtained from 20 supervisors and 50 repairmen, representing four MOSs (63H, 63G, 52D, 45K) and three companies.

Pearson product-moment correlations were calculated to determine the relationship between MPS skill indices and performance ratings; for each rater, a correlation was obtained between the performance ratings he assigned to repairmen and those repairmen's skill indices. While only a handful (16%) of these correlations were found to be significant, the majority (86%) of them were positive. In all, these results suggest a small yet positive relationship between MPS skill indices and performance ratings.

Effects of MPS on Training/Job Assignment

Structured interviews were conducted with shop supervisors, team leaders, and selected repairmen to determine what effects MPS feedback has on maintenance training, job assignment, and morale or motivation.

Although attitudes toward MPS as a training information source and guide were generally positive, the actual use of MPS feedback for training purposes was varied and apparently related to two factors, supervisory level and MOS density: Warrant officers, for example, use MPS feedback more often than team leaders, and MPS reports play a greater role in larger maintenance sections that in smaller ones.

MPS skill training reports were used for training purposes in the following ways: (1) to assign jobs to work teams, i.e, a team which had done some job least often would be assigned that job over other teams, (2) to temporarily shift a repairman from one team to another in order to gain experience on some job, (3) to serve as a memory refresher about which repairmen require training on critical skills, and (4) to log entries into job books.

Interviews with repairmen revealed that individual skill history reports had little effect on morale or motivation for two reaons. First, the skill history reports are distributed infrequently (every six weeks) and second, the reports list which jobs have been performed but not the quality of performance.

Discussion

The primary purpose of MPS is to improve the conduct and quality of on-the-job maintenance training. Toward this end, and as a prototype system, MPS appears to have worked well. It has been shown to be a system in which daily maintenance activities are monitored with relative ease and accuracy. Moreover, MPS provides unique maintenance training information and guidance which is used for training purposes by shop personnel.

The results of the evaluation reported here and other observations made during the past year suggest several improvements or modifications of MPS. For example, modification of the MPS computer program would allow a calculation of direct man-hours from information on MPS Form 2, thus increasing the accuracy and reliability of the man-hour availability and use reports (and eliminating all or at least part of MPS Form 3). Another modification would be to increase the frequency of distribution of individual skill history reports, and hence increase the potential this feedback has for motivation and morale. One other important improvement of MPS now under development is a method of reflecting the proficiency with which tasks are performed; currently, two repairmen will receive the same amount of credit for completing a task regardless of the quality of their work.

ARI is now working toward a link between MPS and interactive videodisc instruction. A scenario for the near future is this: After signing on at a computer terminal, a mechanic is presented with a summary of his current work and training history. The computer then identifies the mechanic's training deficiencies, and directs him toward a specific videodisc lesson available for use on that terminal. The computer monitors the progress of the mechanic as he completes the lesson, and updates his training experience accordingly.

Finally, it should be noted that a maintenance system like MPS on have an impact on more than the unit in which it functions. Specifically, maintenance performance information can be fed back to the institutional training base for use in curriculum design, and it can provide high quality reference data to the equipment design process.

The Strategic Weapon System
Training Program

Part I - Description

Frank B. Braun
DATA-DESIGN LABORATORIES

ABSTRACT

This paper (Part I of a two part paper) describes the Strategic Weapon System Training Program (SWSTP) used in the Navy's POSEIDON and TRIDENT Submarine Force.

The goal of the SWSTP is to provide current, comprehensive, job-related training for Strategic Weapon Systems personnel in support of maintaining the SWS in a high state of operational readiness. The POSEIDON and TRIDENT submarine systems are required to operate in an environment where the ships must be completely self-sufficient, yet maintain their weapon system in a constant "up" condition. This requirement necessitates trained technicians who can operate and maintain their equipment without outside help.

The SWSTP, developed and refined over the past 10 years, has five major elements: Personnel Performance Profiles, Training Path Systems, Curricula, Personnel Qualifications Guides, and the Personnel and Training Evaluation Program. These elements and the supporting organization are described.

BACKGROUND

Training for the POLARIS Fleet Ballistic Missile (FBM) System was reviewed in the late 1960's and was found deficient in several respects. The actual training requirements were confusing, training pipelines were excessively long, overtraining was common, trained personnel often were unfamiliar with current technical publications, and no adequate method existed for evaluating training effectiveness.

To correct these problems, the Chief of Naval Operations directed the establishment of the FBM Weapon System Training Program and assigned specific responsibilities for its implementation and administration. The advent of the POSEIDON and TRIDENT systems required further refinements to the training program and the name was changed to the Strategic Weapon System Training Program (SWSTP). The Strategic Systems Project Office (SSPO) was designated to implement and provide overall technical control of the program.

PROGRAM ELEMENTS

The SWSTP is a systems approach to training composed of five major elements. The Personnel Performance Profiles (PPPs) are comprehensive, minimum requirements listings of the knowledges and skills required to operate and maintain a system, subsystem or equipment. The PPPs are essentially the result of hardware-oriented task analyses and are prepared using current information from approved engineering drawings, technical manuals, training literature, contractor source information and promulgated operational, maintenance, safety and training doctrine.

The PPPs are in the form of numbered tables corresponding to the system, subsystem or equipment configuration of the SSBN Weapon System. For example, PPP table 1500 contains items which list knowledge and skills associated with a Navigation Subsystem. The individual system, subsystem or equipment table consists of two parts: (1) knowledge items representing the theory, characteristics, functional operation and procedures involved in the operation and maintenance of the system, subsystem or equipment; and (2) skill items representing the abilities required to perform operation and maintenance based on acquired knowledge.

PPP items are written to encompass all depths (levels) of coverage. No attempt is made to list every detail of every type of knowledge or skill required. It is the responsibility of the PPP user to ensure comprehensive detail coverage in material derived from the PPPs.

The Training Path System (TPS) assigns the knowledge and skill items of the PPPs to specific Navy personnel in a logical order and to a defined depth of knowledge and level of skill. For example the TPS indicates which PPP table 1500 items are required in a particular classification-e.g., SINS Technician-and identifies the courses which must be taken to acquire those skills and knowledge. The logical order of those courses forms a training path. The TPS consists of three components: (1) Training Objective Statements (TOS) which define the level of training coverage for knowledge and skill items of the PPP for specific Navy personnel; (2) Training Path Charts (TPC) which present graphically the PPP items and TOS levels required for each course and the logical sequence of courses for specific personnel; and (3) Training Level Assignments (TLA) which assign the level of training to particular PPP items for specific personnel and also identify the type of training (i.e., Background, Replacement, Advanced, and Onboard). In conjunction with the PPPs, the TPS constitutes an effective management tool for coordinating the development of training materials and for ensuring adequate and uniform coverage of subject matter, with a minimum of duplication, in training courses and materials.

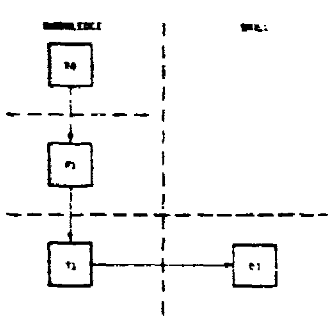
Alphanumeric codes are used in the TOS to identify knowledge and skill categories and depth of coverage. The alphabetic portion of the code indicates the knowledge and skill category:

Knowledge:
F - Familiarization Theory
T - Theory

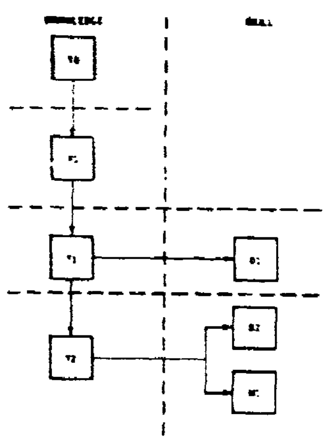
Skill:
O - Operation
P - Preventive Maintenance
C - Corrective Maintenance
M - Maintenance

The numeric portion of a code will be 1, 2, or 3 (3 is the highest level reached in any knowledge or skill). TOS "TO" is an exception and refers to the background knowledge and skills usually taught in "A" School. An example of a training level is "C2," which refers to advanced corrective maintenance skills used in undocumented fault isolation procedures performed with test equipment.

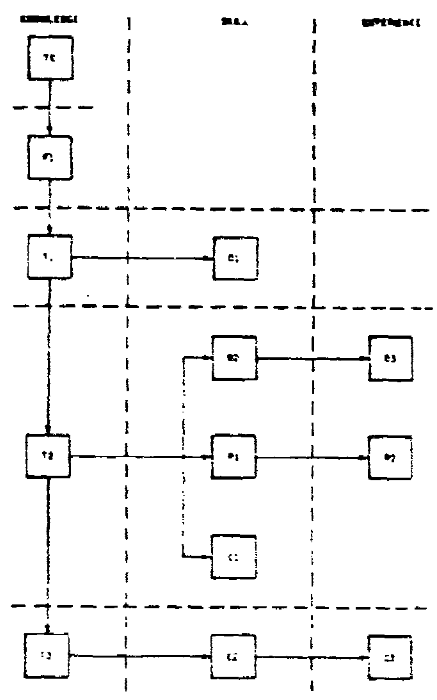
The TOS for knowledge and skills are grouped into functional task sets according to three categories of personnel: CO/XO (Coordinate), officer (except CO/XO) (Direct), and technicians (Perform). Related statements should be read as a group to more clearly understand each of the statements, the increasing coverage, and the differentiation between statements. To provide a better understanding of the TOS, the following diagrams illustrate the increases in level, the relationships between TOS levels, and the paths of training.



Coordinate Task TOS Set



Direct Task TOS Set



Perform Task TOS Set

Note that in each diagram, a knowledge statement (TOS level, T1, T2, or T3) is followed by one or more skill statements before another knowledge statement is required. In other words, each knowledge supports one or more skills: T1 supports O1; T2 supports O2, O3, P1, P2, and C1 (O2 and M1 for officers); and T3 supports C2 and C3. In addition, it should be understood that T1 is supportive to T2 and T2 is supportive to T3, although this is not explicitly indicated in the diagram. Construction of the diagrams in this manner allows diagrams to be modified to reflect specific training requirements for particular personnel.

The entire diagram or a modified diagram would be applicable to each PPP table indicated as being part of the training coverage for particular personnel. The entire diagram reflects the training requirements for personnel associated with all operation, preventive maintenance, and corrective maintenance.

A TPC, the second component of the TPS, consists of four parts: (1) Table Assignment Matrix, (2) PPP Table Index, (3) Table Assignment Chart, and (4) Curricula Index. A separate TPC is presented for each group of POLARIS, POSEIDON, and TRIDENT COs/XOs, officers and SWS technicians.

The Table Assignment Matrix indicates the particular PPP coverage (by PPP table number) required for training the particular officer group or type of technician. The table assignment matrix also indicates the TOS levels of training required for each listed PPP table.

The PPP table index provides the PPP table name corresponding to each table number. The Table Assignment Chart provides a graphical presentation of a complete training pipeline for a category of personnel. The chart identifies the areas of knowledge and skills that personnel of a specific category are expected to acquire during their training life. If required, the chart may also identify entry or exit points to or from another training path. The TPC indicates three categories of training: Background, Replacement/Conversion, and Advanced/Onboard Training. The Background, Replacement, Conversion, and Advanced training is conducted in formal courses of study (curricula). The table assignment chart specifies the PPP coverage provided by each curriculum. Onboad training requirements are also specified by blocks indicating particular PPP coverage.

The curricula index presents the titles of all formal curricula associated with a particular table assignment chart. The required TOS levels to be covered by each curriculum are also listed.

The TLA, the third element of the TPS, indicates those required PPP items (knowledge and skills) and their assigned levels of training for each group of Strategic Weapon System officers and each Strategic Weapon System technician. The TLA further indicates the relationship of the PPP table items to the TOS codes. The PPP table knowledge categories 1-1 through 1-6 are directly related to the TOS codes F and T, and the skill categories 2-1 through 2-3 are directly related to the codes, O, P, C, and M.

The groups of officers and technicians are the same as those described for the TPC; therefore, for each TPC there is a corresponding TLA.

Curricula is the third major element of the SWSTP. The curricula are designated either formal or informal and are designed to accurately reflect the training requirements identified in the TPS. Formal curricula are used in training facilities ashore to provide background, replacement, conversion, or advanced training. They include Instructor Guides, Trainee Guides, and written and performance tests.

The Instructor Guide is the primary element of every curriculum. It establishes the detailed course learning objectives, sequences the presentation of information, and programs the use of all other curriculum elements.

Informal Training is received on board submarines and tenders or ashore outside the formal classroom. Informal training is accomplished either through work under the supervision of more experienced SWS personnel or through use of instructional media such as Self-Study Workbooks, Lecture Guides (programmed slide presentations), and videotapes. Self-Study Workbooks are specifically designed for use on a self-paced basis and cover topics at system, subsystem, and equipment levels. Informal training is also used to train individuals on new equipment or equipment altered from its original configuration.

Curricula materials are acquired and maintained through a comprehensive system of stage submittals and review which are designed to ensure the use of curricula which reflect current training requirements on the theory, maintenance, and operation of the applicable system, subsystem or equipment, and in the proper use of associated documentation and procedures.

The Personnel Qualification Guides (PQGs) are promulgated by the Submarine Force Commanders. The PQGs identify specific knowledge and skill requirements or standards that must be met by personnel to "qualify" or "requalify" for various watchstations on board a submarine or tender. These qualifications differ from training in that they require a specific demonstration of ability using the actual equipment in its operational environment after appropriate training has been completed. Replacement training provides the necessary knowledge and skills to pursue those qualifications requiring the least experience.

The Personnel and Training Evaluation Program (PTEP) is the final element of the SWS Training Program and is the quality assurance element. The PTEP measures, evaluates, and reports on the effectiveness of the program. It is designed to assist overall training management by providing a capability for training monitoring, evaluation, feedback, and improvements. PTEP accomplishes its objectives by means of personnel testing, collection of test and nontest data, evaluation, and reporting. Both test and nontest data are evaluated to determine trends, identify deficiencies, and formulate recommendations for corrective action. Evaluation results are promulgated in various types of reports which are distributed to assist commands in increasing personnel proficiency and in implementing improvements to the SWSTP.

Tests are administered periodically to personnel from the time they enter training until they leave the system. The PTEP utilizes standardized tests referred to as System Achievement Tests (SATs) and Course Achievement Tests (CATs) for individual and training program assessments.

SATs are administered periodically to all technicians in the SWSTP to measure overall system knowledge and skill levels. They are also given at the completion of replacement training and are based on the PPPs and TPS for each technician group. The primary purpose of SATs is to evaluate the training system. Their secondary role is to assess an individual's knowledge and skill and those of his crew.

CATs are administered at the end of each major instructional phase and at the completion of a formal course. The test content is defined by the knowledge and skill requirements of the PPPs and TPS as contained in the Instructor Guide (IG). Reports of tests are provided to training commands to assist them in improving instruction.

PTEP Data Collection (nontest) is the process of collecting and maintaining information, other than test results, to provide the basis for test validity studies, group performance evaluations, and other evaluations of the SWS Training Program elements. Nontest data include personal history data, group performance data, the status of training facilities/hardware/documentation, and other pertinent statistics.

The PTEP evaluation component utilizes all accumulated data to measure the effectiveness of the training program. Analyses are conducted to measure individual and crew knowledge and skill levels, curriculum adequacy, and hardware reliability. The analyses are designed to identify and verify deficiencies within the training program and formulate recommended corrective actions.

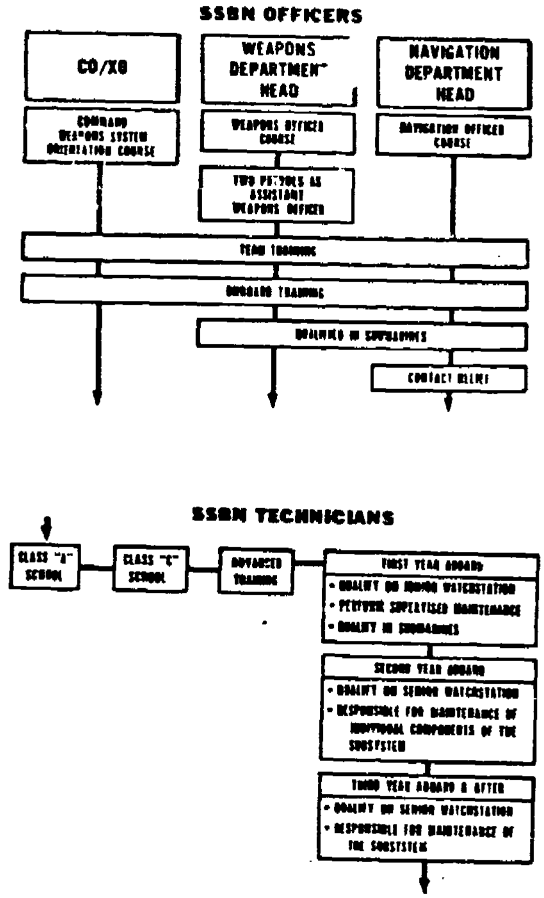
Evaluations are normally initiated on the basis of indications from test results. In certain circumstances, specific evaluations may be initiated as a result of tactical hardware performance, reporting requirements, particular inquiries, or changes in personnel requirements, tactical hardware, training facility/hardware documentation, etc.

PTEP provides a variety of external reports based upon the analysis of test and nontest data. These reports are designed specifically to assist commands in assessing and improving personnel proficiency and the SWSTP. The most widely used PTEP reports are: (1) Group System Achievement Test/Group Course Achievement Test Reports which present a "quick-look" analysis of the results of personnel testing. They include the current test scores for each personnel category, individual scores (for both the overall test and for individual knowledge and skill areas within the test), and an analysis of group performance in each area. (2) Test Evaluation Reports (TERs) which contain the results of a statistical analysis of each test item, each skill test exercise, and each knowledge and skill area. Data from TERs provide information about tests, in general, and test items, in particular that is valuable in the design and development of replacement tests. (3) The PTEP Status Report is a semi-annual summary of PTEP testing and evaluation results. It is distributed to activities which have an interest in the current status of the training program but which are not directly concerned with program management. (4) Evaluation Reports and Special Investigations which are developed when a training deficiency is identified. They contain information related to the deficiency, provide pertinent background data, and recommend corrective action.

SWS TRAINING SEQUENCES

Training sequences are established by manpower resource considerations and system requirements. The levels of training established by the TPS for F1, T0, T1, M1, C1, O1, P1, T2, and O2 are normally accomplished at the Replacement Training sites. All other levels of training are normally accomplished in the formal school environment at the Advanced Training sites and/or the On-Board Training (OBT) environment for SSBN personnel. All Training established by the TPS for SWS Tender personnel will normally be accomplished at the Replacement Training site. Tender training which cannot be conducted in replacement training will be accomplished in the SWS Tender OBT environment. Typical training sequences for submarine officers and technicians are depicted below.

TRAINING SEQUENCES



SWSTP ORGANIZATION

The SWS Training Program organization is structured to carry out the policy objectives of the program. The training program organization is depicted in the organization chart on Figure (1). The CNO provides overall policy guidance, approves changes and establishes requirements for training and evaluation. The Fleet CINCs, Chief of Naval Material, and Chief of Naval Education and Training implement the SWSTP through their respective organizations. Chief of Naval Technical Training establishes formal training courses to support the requirements of the TPS and provides, in conjunction with SSPO, the specifications and procedures for planning, developing, implementing and maintaining the programs training material. SSPO provides and maintains current the program elements for the SWSTP and provides technical coordination and support to the Central Test Site for PTEP (CTS). CTS conducts the evaluation tests through CTS Detachments at the SSBN off-crew sites. CTS prepares the tests, analyzes and evaluates the data, prepares reports and recommendations for SWSTP improvements.

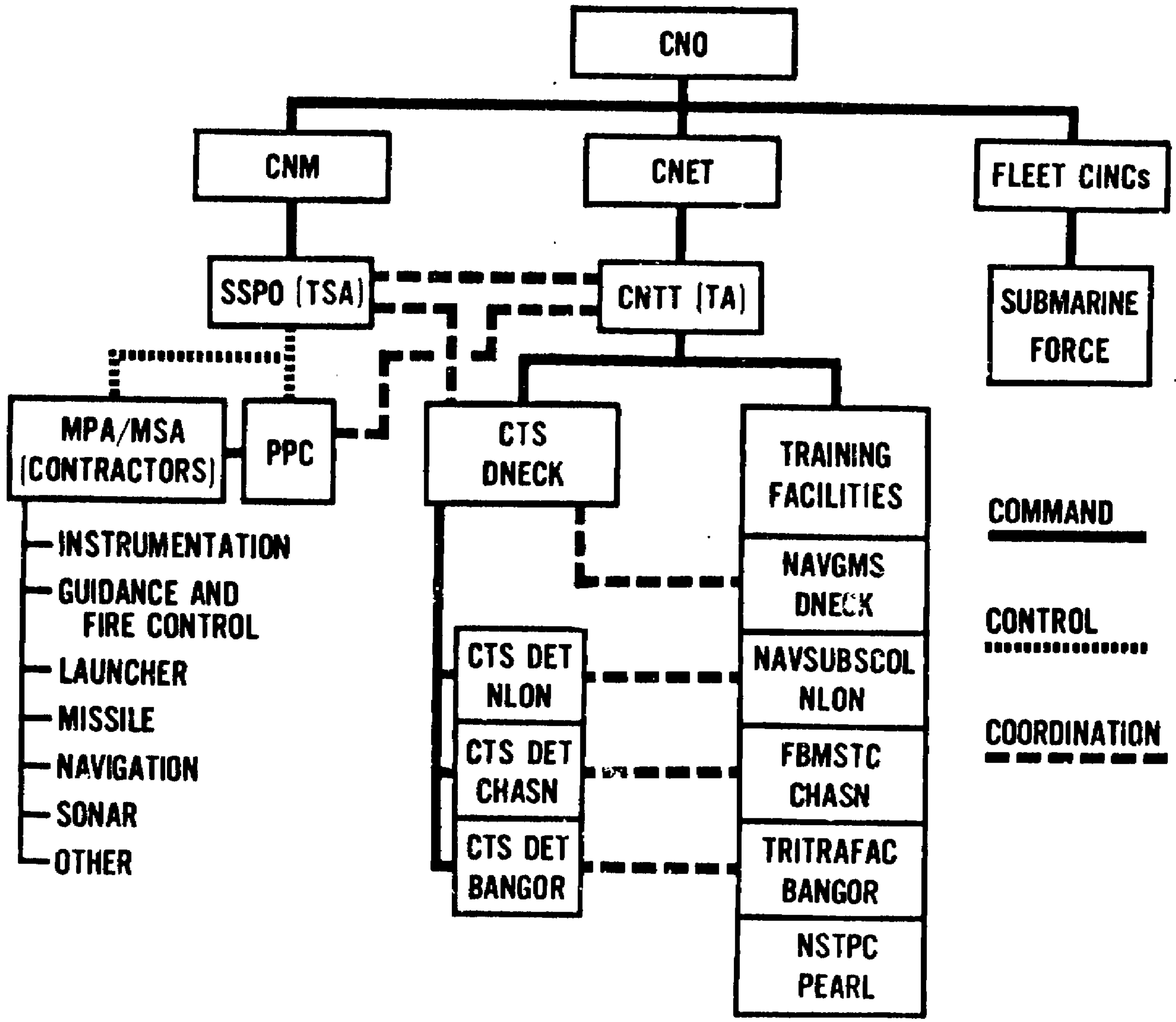

CNO
CNM
CNET
FLEET CINCs
SSPO (TSA)
CNTT (TA)
SUBMARINE FORCE
MPA/MSA (CONTRACTORS)
PPC
CTS DNECK
TRAINING FACILITIES
INSTRUMENTATION
GUIDANCE AND FIRE CONTROL
LAUNCHER
MISSILE
NAVIGATION
SONAR
OTHER
NAVGMS DNECK
NAVSUBSCOL NLON
FBMSTC CHASN
TRITRAFAC BANGOR
NSTPC PEARL
CTS DET NLON
CTS DET CHASN
CTS DET BANGOR
COMMAND
CONTROL
COORDINATION

Figure 1

The Submarine Force Commanders develop and promulgate the SWS personnel Qualification Guides and provide requirements and recommendations for OBT materials to SSPO.

The SWS Training Facilities (TF) conduct formal courses and act as Material Preparing Activities (MPA) and as Material Support Activities (MSA) for the SWSTP Materials Management task.

SWSTP MATERIALS MANAGEMENT

The SWS Training Program materials include management materials, curriculum materials, and instructional media materials. The SSBN Weapon System Training Materials Management Plan is designed to provide timely, adequate, and accurate training materials for the SWSTP. The organizations shown in Figure (1) perform vital functions in the development and support of SWSTP materials in the following roles:

The Chief of Naval Technical Training (CNTECHTRA) as the Training Agency (TA) for the SWSTP exercises command of and provides support to the formal training effort.

SSPO as the Training Support Agency (TSA) is responsible for supporting the TA by providing material and other forms of support within its cognizance. The Personnel Program Coordinator (PPC) is the organization responsible for planning, designing, and providing a fully operational personnel subsystem under direction of the TSA. The PPC provides coordination and review of training material development and support as directed by the TSA. This role is performed by a non-hardware contractor for the SWSTP.

A Materials Change Activity (MCA) is an organization that develops changes to training materials, as directed by the TA or TSA. MCAs for the SWS Training Program are either contractors or Navy activities (e.g., TFs). Materials Preparing Activities (MPAs) develop, revise, and/or produce training materials as directed by the TSA or TA. MPAs in the SWSTP are either contractors or Navy activities (e.g., TFs). Materials Support Activities (MSAs) are responsible for training materials surveillance as directed by the TSA or TA. MSAs in the SWS Training Program are either contractors or Navy activities (e.g., TFs).

SSPO, in its role as the TSA, has developed a "family" of contractors to support the SWSTP. This group of contractors, both hardware and engineering services, have developed an expertise over many years in the program which allows them to respond quickly and comprehensively to new or changed program requirements.

The training materials management program consists of the techniques and procedures for acquiring and supporting training materials. The development and support of each of the three types of training materials are discussed in the following paragraphs.

Management materials consist of documents which contain the PPPs and the TPS and program management documents which contain specifications and procedures. PPPs and TPS are developed by contractor and training facility MPAs, coordinated and reviewed by the PPC, and approved by the TSA and TA. Management procedures and materials specification documents are developed and supported by the PPC, under the direction of the TSA, reviewed by program participants, and approved by the TSA.

Curriculum materials consist of Instructor Guides and Trainee Guides or instruction sheets. Curriculum materials are developed by contractor MPAs or training facility MPAs, coordinated and reviewed by the PPC, and reviewed and approved by the TA and TSA.

Instructional media materials are slides, overhead transparencies, Self-Study Workbooks, programmed super 8mm presentations, Exercise Controller Guides, Administrator's Guides, and Lecture Guides. These materials are developed largely by contractor MPAs, but TFs may also produce transparencies and other materials. The PPC performs coordination and review functions with regard to media materials development and support. If the materials are developed for use in the formal environment, they must be approved by the TA with TSA concurrence. If they are developed for use in the informal environment, they are approved by the TSA only. Since most instructional media materials are changed by revision only, the change procedure is identical to the development procedure.

All training materials are maintained current and accurate by surveillance and change efforts. Surveillance is conducted by MSAs to detect those changes in documentation, equipment, or procedures that impact the training materials. When the requirement for a change is identified by a contractor MSA, a TMCR (Training Materials Change Recommendation) is prepared and forwarded to the PPC. Deficiencies in training materials identified by Navy activities, such as training facility MSAs, will be reported by the use of a TFR (Trouble and Failure Report) or a letter to the TSA. In all cases, the surveillance and reporting of training materials deficiencies are coordinated by the PPC.

Once the need for a change has been identified and reported, an interim change, change, or revision may be initiated. Interim changes and changes are prepared by MCAs and revisions are prepared by MPAs. The type of change required depends on the magnitude and urgency of the change.

SUMMARY

The operational readiness of the fleet is highly dependent upon well-trained officer and enlisted personnel. As new weapons and delivery systems become more sophisticated and complex, the need for adequate and cost-effective training programs becomes more acute. An effective training program requires accurate identification of training requirements, uniform training for personnel at all sites, and a timely method for updating materials. The SWS Training Program, utilizing a systems approach and constant surveillance, meets this need for the Weapons and Navigation personnel on our strategic submarines.

Braun, Henry I. & Jones, Douglas H., Educational Testing Service, Princeton, New Jersey. (Thurs. P.M.)

AD P001295

Empirical Bayes Assessment of Differential Validity

An important problem in the assessment of test validity arises when the population under consideration consists of several distinct subgroups distributed in varying proportions across a number of training sites or programs. Separate prediction equations for each group at each site may be required, particularly if there is evidence of differential validity. The authors describe and illustrate a technique, based on the empirical Bayes paradigm, which provides estimates of the different prediction equations even when the members of a particular group are sparsely scattered across sites so that unique least squares estimates are not available. There is no limit to the number of groups, sites, or explanatory variables that can be considered.

Unbiased Predictions in Sparse Data Problems

Henry I. Braun and Douglas H. Jones
Educational Testing Service
Princeton, New Jersey

The Graduate Management Admissions Test (GMAT) plays an important role in the admissions decisions made at many graduate management schools. A candidate's GMAT score is often combined with his/her undergraduate grade point average in a multiple regression equation to obtain a predicted first year average. This predicted score is but one component in the evaluation of the candidate's suitability for study. It is extremely important that these predictions be fair; that is, that no discernible group of candidates have their predicted scores consistently higher or lower than their achieved scores. This notion of "fairness to the group" is only one way of posing the problem of prediction bias. Swinton (1981) provides an extensive discussion of other possible formulations as well as a lucid survey of the literature on the subject.

One purpose of this study has been to assess the quality of the predictions of first year averages in graduate management school when a single prediction equation is employed for both black students and white students at a given school. Our analysis shows that prediction bias does exist in the sense that a single equation yields predictions for minority students with low predictor scores which, on average, tend to be somewhat higher than the criterion scores they actually achieve. This finding is consistent with much of the previous research in this area (Rolph, et al., 1978; Swinton, 1981).

A second goal has been to develop a new bias-free prediction system based on GMAT scores and undergraduate grade point average (UGPA). The proposed system explicitly employs race as a predictor. Although there are no statistical difficulties, the ethical and perhaps the legal propriety of this approach is not well established.

Because of the relatively few black registrants in our data base, it would be virtually impossible for the conventional least squares procedures to provide parameter estimates for the type of model we propose. We have employed, therefore, the empirical Bayes methodology (Maritz, 1970). Although several technical difficulties have had to be overcome, we have succeeded in obtaining useful parameter estimates that should prove fairly stable over time.

The Data and Evidence for Prediction Bias

This study is based on data collected at fifty-nine business schools in the academic years 1978-79 and 1979-80. The final data base consists of about 8500 records of which approximately four percent represent those of black students.

One approach to the assessment of prediction bias is to estimate separate prediction lines for majority and minority students. However, as in most studies of this nature, there are three obstacles to such a simple analysis. The first is the complicated nature of the selection process (restriction of range on measured predictors as well as possible selection using unmeasured predictors). A second difficulty is the questionable comparability of the criterion across schools. Finally, small minority sample sizes complicate the estimation process.

The scarcity of minority data is illustrated by the fact that of the 59 schools, 43 have <u>five</u> or fewer blacks enrolled per year of participation in the GMAT Validity Study Service. These 43 schools enroll about 20 percent of the minority students. It is, of course, virtually impossible to estimate an accurate prediction line for a school having only five data points, especially with three predictor variables. Even in the five schools with 20 or more black students, restriction of range renders it difficult to appraise the nature of the prediction bias separately within each school.

Inasmuch as there are only a few schools with moderate numbers of black students, the simplest way to determine the existence of predictive bias is to estimate a single multiple regression equation at each school for all the students. To make the outcome variable comparable across schools, the criterion or dependent variable chosen is the Z-score equivalent of the student's first-year average. The Z-score equivalent is the number of within-school standard deviations above or below the within-school mean that a particular score falls.

Once the estimated combined prediction line is obtained by ordinary least squares, a predicted Z-score FYA is computed for each black student and is subtracted from that student's actual Z-score FYA to obtain a residual. For conciseness this is called black residual. Within each school, the black residuals are averaged to obtain a mean black residual. The mean black residual indicates if, on the average, the combined prediction line over or under predicts for the black students in that school.

The results of the computations yield a preponderance of negative mean black residuals, indicating that the single least squares prediction line tends to overpredict FYA for black students at a large majority of schools. Further investigation reveals that this apparent bias is not related to the fact that minority and majority predictor scores are concentrated in different regions. More specifically, the least squares line leaves small, symmetrically distributed residuals for white students whose predictor scores are located around the centroid of black predictor scores.

The Empirical Bayes Method

In order to overcome the difficulties mentioned in the previous section, particularly the problem of small sample sizes, previous studies have employed various methods of pooling the data. These have generally consisted of a single regression equation with several parameters common to all schools and a limited number of parameters for individual schools (Boldt, 1975; Tucker, 1963 and Potthoff, 1964). In order to produce stable prediction systems when many small schools are involved, the individual school parameters are usually suppressed and a single equation is employed for all (Wilson, 1979).

An entirely different approach, based on Bayesian principles, permits each school to have its own prediction equation, even when sample sizes are small. One methodology has been developed and implemented by Lindley (1970) and Novick, et al., (1972). However, the desired estimates are not easily obtained and their numerical values are quite sensitive to the computational routines employed. Furthermore, these estimates can be substantially improved when many parameters are involved.

Motivated by Bayesian ideas, statisticians have also developed the so-called empirical Bayes techniques (Robbins, 1951; Efron and Morris, 1973) which require reduced specification of prior information — critical to Bayesian techniques. Instead, the data is allowed to determine the probable shape of the prior distribution.

The principal benefit of empirical Bayes, relative to traditional methods, is that it yields parameter estimates which are more stable through time and better predictors for future data (Rubin, 1980). More importantly, in this study we have exploited another aspect of the methodology that permits the estimation of separate prediction equations for whites and blacks in each school, even where the number of blacks is too small to permit the classical estimator to be uniquely defined. Moreover, it employs whatever information is available for black students in a meaningful and plausible way. The following sections will provide technical details and numerical results.

The Empirical Bayes Model Prediction

The description of the empirical Bayes model involves two components: (i) the conditional distribution of the observed criterion scores given the unobservable parameters and (2) the distribution of the unobservable parameters themselves. Before proceeding, we require some notation. As before, i indexes schools and j indexes students within schools.

Let

Y'_{ij} = observed first year average (FYA),

Y'_i = average FYA in school,

N_i = sample size in school i,

$Y_{ij} = (Y'_{ij} - Y'_i) \big/ [\Sigma_j (Y'_{ij} - Y'_i)^2/N_i]^{\frac{1}{2}}$ = standardized FYA,

V_{ij} = score on GMAT (verbal),

Q_{ij} = score on GMAT (quantitative),

U_{ij} = undergraduate grade point average (UGPA),

$I_{ij} = \begin{cases} 1 & \text{, if student is black} \\ 0 & \text{, if student is white.} \end{cases}$

The first component of the model can be described by saying that given $B_{0i}, B_{1i}, \ldots, B_{7i}$, the observed standardized FYA is generated according to the regression:

$$(1) \qquad Y_{ij} = B_{0i} + B_{1i}V_{ij} + B_{2i}Q_{ij} + B_{3i}U_{ij} + I_{ij}\left\{B_{4i} + B_{5i}V_{ij} + B_{6i}Q_{ij} + B_{7i}U_{ij}\right\} + \varepsilon_{ij}$$

where ε_{ij} are independent normally distributed errors with mean zero and unknown variance σ_i^2.

In this model, if the jth student is black, all eight terms of equation (1) are employed in predicting his standardized score; if he is white, only the first four terms are used, since $I_{ij} = 0$ in that case. In other words, the prediction equation for blacks is obtained by appending to that for whites an adjustment to the intercept and coefficients of the three predictors. For example, the coefficient of UGPA in the prediction equation for blacks is $(B_{3i} + B_{7i})$, whereas for whites it is simply B_{3i}.

Geometrically, equation (1) is equivalent to requiring two different prediction planes in each school: one for whites and one for blacks. Until this point, there has been no mention of parameter estimation, and either least squares, empirical Bayes or some other methodology could be employed. However, with the small minority sample sizes available here, least squares is entirely impractical. Hence, we augment (1) with some assumptions which place the model in empirical Bayes framework.

The second component of the model specifies that the vectors $\mu_i = (B_{0i}, B_{1i}, \ldots, B_{7i})$ $(i = 1, \ldots, 59)$ are a random sample from a multivariate normal distribution, $N(\mu^*, \Sigma^*)$. The parameters μ^* and Σ^* are not specified, but rather are estimated from the data under the above distributional assumptions which seem reasonable in this context.

Even with the normality assumptions, maximum likelihood estimates of μ^*, Σ^*, and $\mu_i (i = 1, \ldots, 59)$ can be obtained in closed form. An E-M algorithm (Dempster, Laird, Rubin (1977)) provides a relatively simple iterative scheme for finding the maximum likelihood estimates, as well as estimates of their variability. The latter are required to assess the significance of the differences between regression coefficients for blacks and whites.

Results and Conclusions

Our methods and analysis are designed to answer two questions:

(1) Under the two-plane model, are the differences between the prediction planes for white and black students statistically significant and of practical importance?

(2) If practical differences do exist, are the predictions for black students superior to those obtained under the classic model?

The first question may be addressed in two ways. One may compare individual regression coefficients for black and white students and assess the magnitude of their differences. This yields a substantial number of statistically significant results at the 5 percent level. However, instability of individual coefficients suggests that the second method, described below, is preferable. Here one considers the differences in predictions between the two models at a number of preselected points in the predictor space. We have chosen to focus on that point with coordinates equal to the mean black student score on each predictor at the particular school. In effect we construct for each school a typical set of predictor scores and ask by how much the empirical Bayes predictors of <u>average</u> white student performance and <u>average</u> black student performance differ at this point.

The results are quite definitive. More than half the differences are greater than a quarter of a standard deviation (of the score distribution in the school). About a quarter of the differences are statistically significant at the 5 percent level. The answer to question (1) appears to be in the affirmative.

Turning to the second question, we find that overall the black residuals from the empirical Bayes model are symmetrically distributed and smaller in magnitude than those from the classical model. The only exception is the relatively small group of black students with rather high predictor scores. They are somewhat better predicted by the least squares line. This finding is somewhat at odds with what the "regression to the mean" theory would suggest and deserves further study.

We have also compared how well the predictions of the two models correlated with the scores actually obtained by black students. The empirical Bayes model proved superior in this comparison as well, although the method is not meant to find the best fit for the particular data at hand. In fact, further analysis shows the improvement is due to adoption of the two-plane strategy rather than empirical Bayes methodology per se. Of course, in this setting the two-plane model can only be reasonably fit by using something akin to Bayesian methods.

Although a number of technical and substantive issues remain unresolved, we believe that the use of empirical Bayes methods enables the analyst to successfully construct fair prediction models in such sparse data problems. Those interested in the technical aspects of this approach should consult Braun and Jones (1981).

References

Boldt, R. R. Comparison of a Bayesian and a least squares method of educational prediction. GRE Board Professional Report (GREB No. 70-3p). Princeton, N.J.: Educational Testing Service, 1975.

Braun, H. I., & Jones, D. H. The Graduate Management Admissions Test Prediction Bias Study. Final report to the Graduate Management Admissions Council. Princeton, N.J.: Educational Testing Service, 1981.

Dempster, A. P., Laird, N. M., & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. The Journal of the Royal Statistical Society B, 1977, 39(1), 1-38.

Efron, B., & Morris, C. Combining possibly related estimation problems (with discussion). The Journal of the Royal Statistical Society B, 1973, 35(3), 311-319.

Lindley, D. V. A Bayesian solution for some educational prediction problems, 3 (Research Bulletin 70-33). Princeton, N.J.: Educational Testing Service, 1970.

Maritz, J. S. Empirical Bayes Methods. London: Methuen, 1970.

Novick, M. R., Jackson, P. H., Thayer, D. T., & Cole, N. S. Estimating multiple regressions in m-groups: A cross-validation study. British Journal of Mathematical and Statistical Psychology, 1972, 25, 33-50.

Potthoff, R. F. The prediction of college grades from College Board scores and high school grades. Institute of Statistics Mimeo Series 419. Chapel Hill: University of North Carolina, 1964.

Robbins, H. An empirical Bayes' approach to statistics. Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability. Berkeley and Los Angeles: University of California Press, 1951, 131-148.

Rolph, J. E., Williams, A. P., & Laniear, A. L. Predicting minority and majority medical student performance on the National Board exams (R-2029-HEW). Santa Monica, CA.: Rand Corporation, 1978.

Rubin, D. B. Using empirical Bayes techniques in the law school validity studies. Journal of the American Statistical Association, 1980, 75, 801-827.

Swinton, S. Bias in graduate admissions tests. Graduate Management Admissions Test Research Report, Princeton, N.J.: Educational Testing Service, 1981.

Tucker, L. R. Formal models for a central prediction system. Psychometric Monograph No. 10, Richmond, Va.: William Byrd Press, 1963.

Wilson, K. M. The Graduate Record Examination Cooperative Validity Studies Project. GRE Board Research Report (GREB No. 75-8R). Princeton, N.J.: Educational Testing Service 1979.

7

Bridges, Claude F., United States Military Academy, West Point, New York.
(Tues. A.M.)

A New Approach to Developing Test Scales

Development of new scales from test items usually requires several hundred cases in each of the defined groups. Application of multivariate techniques to existing test scales also encounters such problems as multicolinearity, combining ipsative with normative measures and ordinal with equal interval scales, differential prediction and communication of multiple scale scores to untrained users. To meet these problems a method was devised whereby multiple scales from different tests could be treated as items, yielding new scales with high validities (several in the .90's) and normal to negligible shrinkage, even when developed on relatively small groups. The results were well-received by cadets, counselors, and educators, and for two years have been incorporated as a regular part of the cadet academic counseling program. The statistical development procedures and the results are illustrated using the Strong-Campbell Interest Inventory based academic field "Interest Profile Similarity Scales."

A NEW APPROACH TO DEVELOPING TEST SCALES[1]

Mr. Claude F. Bridges, Research Psychologist

Office of the Director of Institutional Research
United States Military Academy
West Point, New York 10996

ABSTRACT

Development of new scales from test items usually requires several hundred cases in each of the defined groups. Application of multivariate techniques to existing test scales also encounters such problems as multicolinearity, combining ipsative with normative measures and ordinal with equal interval scales, differential prediction and communication of multiple scale scores to untrained users. To meet these problems a method was devised whereby multiple scales from different tests could be treated as items, yielding new scales with high validities (several in the .90's) and normal to negligible shrinkage, even when developed on relatively small groups. The results were well-received by cadets, counselors, and educators, and for two years have been incorporated as a regular part of the cadet academic counseling program. The statistical development procedures and the results are illustrated using the Strong-Campbell Interest Inventory based academic field "Interest Profile Similarity Scales."

PREFACE

The former title was "New Scales from Old," It was changed because it seemed to be too "cute" and connoted only one of the possible applications. However it is a succinct presentation of one major application of the concept. Many well-developed standardized tests, such as the Minnesota Multiphasic Personality Inventory and the Strong-Campbell Interest Inventory (SCII), have so many scales that commercial computer scoring services are almost essential, even if the Keys were available. Adding new scales based on new item keys for other criterion groups or for special purposes commonly is impractical, and always requires a large number of cases.

Even when enough cases are available in new criterion groups for use of the usual multivariate techniques, some very difficult problems often are encountered. When independent variables from several sources are to be combined, the different measures may be ipsative or normative, and may have nominative, ordinal, equal interval, and occasionally even ratio-type scales. Multicolinearity of the standard scales commonly results in failure to make full use

---

[1]Any conclusions in this report are not to be construed as official U.S. Military Academy or Department of the Army positions unless so designated by other authorized documents.

of the information available in the data, since only one of several correlated scales may be used. This results in loss of precision, reproducibility, and validity.

If, in effect, scales are treated as test items, multicolinearity of scales could become an asset. When selecting items for keying to form one scale, desired multicolinearity is handled by merely adding together the scores from each item, giving each response the desired weight. This most commonly is a weight of one, but Claudy (1978) recently showed that a biserial weighting procedure is superior. Treating standard scales in much the same way as items, and capitalizing on refinements thus possible, results in a practical method for developing new special purpose scales from the regular scales. This method has now been applied, refined, validated, and cross-validated, three times at USMA--the Academic Years of 1978-79, 1979-80, and 1980-81, and four reports prepared (Bridges: 1979, 1980, and 1981; Shuford: 1980).

METHOD

Overview of New Scale Development

a. Compute, for each variable (such as the standard scale provided by scoring service), the point-biserial correlation with membership vice nonmembership in the criterion group.

b. Select scales with highly significant correlations (usually the 30 to 60 most significant); use the z-transformation value for each correlation as the weight (W) to be assigned to scores on that standard scale (n), and sum their products to obtain the new scale raw score (NSRSC).

$$NSRSC = \Sigma\, W_n X_n$$

c. Compute the raw scores and their mean and standard deviation for the criterion scaling group.

d. Use these to determine the desired New Scale standard score equation,

$$\text{NS Standard Score} = 10^* (\text{NSRSC} - \text{Mean NSRSC})/\text{SD NSRSC}$$

Variables

a. The 176 scales from the Strong-Campbell Interest Inventory (SCII) (Campbell, 1977). These included the Holland Theme Scales, the Basic Interest Scales, the Occupational Scales, the Academic-Orientation Scale, the Introversion-Extroversion Scales, and the administrative scales. The SCII was selected because of its preeminence as the most extensively researched and normed measure of an individual's pattern in various types of interests. This is the basic instrument used in all the projects.

b. The 36 value Scales from the Rokeach Value Survey (Rokeach, 1973). These short, easy to give, and well-researched measures of an individual's systems of Instrumental and Terminal Values, previously had been given to the USMA Classes of 1978 and '79 in connection with other research (Bridges, 1973).

c. High School Rank Score. The scaled score based on cadet's rank at graduation from high school was used as an independent variable in the first project.

d. The cadets' academic averages, by department, on the courses they had taken at USMA during their first three terms (first study only).

e. Each cadet's final cumulative academic average on the courses required for their academic area of concentration (in all three projects), and for their field of study (in the last two projects). In the first project, each of the four Area QPA's at graduation were used also as dependent variables for cadets in that area.

f. A dummy variable for graduation in each area or field. Each cadet was scored "1" if graduating in the given area, or field, and "0" if graduating in any other area or field. Only the four areas of concentration (Applied Science and Engineering; Basic Science; Humanities; and National Security and Public Affairs) were used in the first project. Fields for which there were 25 or more graduating cadets were used also as dummy variables for the other two projects. All of the foreign language fields were also grouped together as another dummy variable. The dummy variables were the dependent variables.

g. Two experimental dummy variables suggested by Biglan's model (1973). For one dummy variable, cadets in either of the established paradigm ("Physical Science") areas received a score of "0." Those in "Nonphysical Science" areas, a score of "1." For the other dummy variable, cadets in the more applied areas (AS & E and NS & PA) were scored "1," those in the "purer," more basic areas (Basic Science and Humanities) were scored "0" (in second study).

Samples

The total samples included all cadets in the respective graduating classes. The numbers varied in the successive analyses due to missing data on one or more of the variables involved.

	Class of 1978	Class of 1979	Class of 1980
Total	844	843	909
Applied Sciences	299	308	312
Basic Sciences	79	90	137
Humanities	95	62	80
NS & PA	335	333	294
Management (Interdisciplinary)	36	50	86

Procedure

a. The AY 78-79 study.

(1) The steps summarized in above overview were used to develop three complete sets of scales to discriminatingly characterize cadets who graduated in each of the four academic areas of concentration from all other cadets. The equations for the SCII Occupation Interest-based scales included weights only for point-biserial correlations significant at the .001 level. A few correlations with lower significance levels were used for the SCII-based General Interest scales and for the Rokeach Value-based scales. The three sets of raw scores were converted to standard scores with a mean of 50 and a standard deviation of 10, for cadets who had graduated in the upper half of the class on their respective Area QPA's. These were called Area Compatibility Scores, though they did not include direct measures of the cadet's differential abilities for the courses in each area.

(2) Using the same procedure, four sets of raw score equations were likewise developed with earned area QPA's as criteria. The set for each area included an interest pattern raw score, a Rokeach values pattern raw score, and an early USMA academic performance raw score. The three area raw scores and high school rank score were combined into four regression equations that yielded the most accurate prediction of each of the four Area QPA's. These were used to produce for each group Area Predicted QPA Scores, scaled to a mean of 50 and a standard deviation of 10 for cadets in the relevant area.

(3) The equations for the new scales developed on portions of the Class of 1978 were used to compute area compatibility scores and four predictive area QPA's for the entire Class of 1978 and, for cross-validation, also for the Class of 1979. Since the compatibility scores were for cadets' guidance in choosing an area of study, distributions of these scores for cadets in each area were more important than correlations as validity information for these scores. These data were tabled for cadets in each half on their Area QPA.

(4) A simplified computer printout, listing the cadet's compatibility scores for each area, was provided to a sample of company academic counselors to obtain experience in their delivery and use.

(5) To investigate the extent to which cadets currently were graduating in areas closely conforming to their pattern of interests, values, and academic abilities, and also to determine the dimensional structure of the areas and its conformance to Biglan's (1973) model, a discriminant analysis was performed with the four area predicted QPA's as independent variables and the four areas as groups.

b. The AY 79-80 Study.

(1) The independent variables were limited to those from the SCII. Cadets graduating in the Classes of 1978 and 1979 were combined. To the dependent variables were added the 21 fields of study having 25 or more cases. In addition, cadets in the seven foreign languages were grouped together.

(2) The 21 field of study, and the four Area of Concentration, Interest Profile Similarity Scales were developed and reported for all cadets in the Class of 1982. The computer produced profile was used as desired by the Academic Counselors during their guidance sessions with their cadets.

(3) The counselor's responses to a questionnaire on all aspects of the Academic Guidance Program were obtained, analyzed, and reported (see Bridges, 1980).

c. The AY 80-81 Study.

(1) The graduating cadets in 1978, 1979, and 1980 were combined by area and field of study. Thus, it was possible to add four more fields of study, and refine the few field of study IPSS scales that, in comparisons with other closely related fields, had been found to yield interpretative anomalies the previous year. Cadets graduating in Physics and those in Computer Science had been found typically to have higher IPSS scores for mathematics than for their own field. Likewise, cadets in Literature typically had higher Foreign Language IPSS scores.

(2) The computer produced IPSS Profile was refined and used by the counselors with cadets in the Class of 1983 in choosing their field of study.

(3) Questionnaire surveys of both counselors and cadets were conducted for further refinements of the materials and information on their use provided in the Dean's Academic Counseling Program.

RESULTS AND DISCUSSION

The FY 78-79 Study

a. Validity. An exceedingly long paper would be required to present all of the relevant useful results. We can only hit some of the highlights. (Those interested in more details will find them in Bridges, 1979.) For guidance in choosing one of the four areas, the primary concern is the rank of each of the four areas relative to each other, rather than their absolute magnitude. Thus, one of the best indications of their validity is the extent to which cadets successful in an area tend to get higher scores for that area than for other areas. Cadets who graduated in each area were divided into two groups on the basis of their Quality Point Average at graduation, based only on the courses directly relevant to their Area. Table 1 presents the mean score made on sixteen scales by cadets who graduated in each half on their area specific QPA. Note that both groups of cadets in each area characteristically had a significantly higher average on their own area guidance scores than on the guidance scores for any of the other areas. However, cadets in the Low-Half Area QPA group for Basic Sciences tended to have about the same or slightly higher scores on the scales for Applied Sciences and Engineering than on the corresponding scales for their own area. Likewise, Low-Half Humanities area cadets had the same or slightly higher NS & PA scores than for Humanities. Incidentally, the regression equations used in developing the predicted Area QPA's indicate that most of the weight in the predictions of Area QPA goes to the cadet's previous academic performance, but the values

Table 1. Mean Area Profile Similarity Scale Standard Scores Made by Class of 1978 Cadets who Graduated in Each Half on their Area-Specific-Courses QPA

Group			Applied Sciences & Engineering Scores			Basic Sciences Scores			National Security & Public Affairs Scores			Humanities Area Scores		
Area	Half	N	AOISS	ARVSS	APQPASS	BOISS	BRVSS	BPQPASS	NOISS	NRVSS	NPQPASS	HOISS	HRVSS	HPQPASS
A	Lo ½	98	46	47	44	40	37	42	37	48	51	33	42	41
A	Hi ½	98	50	50	57	45	40	53	33	43	61	30	40	48
A	Total	196	48	49	51	43	38	47	35	46	56	31	41	45
B	Lo ½	33	47	47	46	45	44	43	34	44	52	33	42	43
B	Hi ½	34	47	42	66	49	50	59	31	45	69	34	43	56
B	Total	67	47	44	56	47	47	50	32	44	61	33	43	50
N	Lo ½	140	30	43	31	26	36	33	50	51	43	44	43	36
N	Hi ½	150	28	43	45	26	38	44	50	50	57	47	43	47
N	Total	290	29	43	38	26	37	38	50	50	50	46	43	41
H	Lo ½	40	28	44	39	27	36	38	49	48	50	48	45	40
H	Hi ½	39	28	41	54	31	39	52	47	50	65	50	50	56
H	Total	79	28	42	46	29	38	45	48	49	60	49	48	50
Total		632	36.6	44.8	45.6	33.8	38.6	43.3	43.3	48.0	54.3	40.3	42.8	44.3
		SD	14.7	10.5	12.8	14.5	11.2	10.2	12.1	9.7	10.6	12.2	10.9	9.2

and interest measures also contributed significantly to increased accuracy of predictions in all except the Basic Science Area. The percent of total prediction of actual Area QPA contributed by each measure is as follows:

Area	USMA Academic Performance	H.S. Rank	SCII Interest	Rokeach Values
Applied Science	71%	3%	14%	12%
Basic Science	88%	10%	-	2%
NSPA	77%	2%	11%	9%
Humanities	65%	-	21%	14%

b. Generalizability. Evidence on the generalizability to other classes of the scaling equations developed on a sample from the Class of 1978 is presented in Table 2. The percent of the total in each group ranged from about 10% for the Humanities Area groups to about 40% for NS & PA area groups. The common correction, dividing the point biserial by the maximum possible for that proportion, increases comparability between areas and between scores. Except for the Area scales based on the Rokeach Values data from administration to the Class of 1979 about three years previously, the scales demonstrated unusual resistance to shrinkage. These concurrent cross-validities are even more impressive when the small number of criterion group cases in some of the developmental analyses for the Basic Sciences and Humanities areas are considered.

c. Dimensional Structure of Curriculum. The Physical Science/Non-Physical Science Score results from combining the cadet's selected compatibility and Predicted Area QPA standard scores into one regression equation that most accurately distinguishes between cadets graduating in the Physical Science areas (Applied and Basic Science areas) and those graduating in the other areas (NSPA and Humanities). The corrected point-biserial correlation between the Physical Science/Non-Physical Science Scores and membership in the Physical Science Group or the Non-Physical Science group was .90 for the sample on which developed, and .85 for the total Class of 1978. When cross-validated by using the same equations for the Class of 1979, the correlation was .83, still remarkably high.

Discriminant analysis, using the four areas as groups, and the four area Predicting QPA Scaled Scores as independent variables, provides confirmation and further structure to be expected if the USMA curriculum conforms to the Biglan Model (Biglan, 1973). Pertinent results follow.

Independent Variables	Standard Discriminant Function Coefficients		
	Func 1	Func 2	Func 3
APQPASS	+2.143	+1.177	+3.407
BPQPASS	+0.449	-0.945	-4.303
NPQPASS	-1.709	+0.803	-0.767
HPQPASS	-0.278	-1.840	+1.731
Canonical Correlation	.70	.31	.13
Relative % Variance	88%	10%	2%

Table 2: The Concurrent Validity and Cross-validity of the Seventeen Scores

Correlations between Each Set of Scores and their Criterion, for All Cadets Having Data on the Corresponding Pair of Variables, in the Total Class of 1978 (Maximum N = 733) and in the Class of 1979 (Maximum N = 741). For the dichotomous dummy variables, the obtained point biserial correlations have been corrected to same base.

	Class of 1978			Class of 1979		
SCORES	N	$r_{pt.\ bis.}$	r.	N	$r_{pt.\ bis.}$	r.
COMPOSITE A/B-N/H	255	(.674)	.853	307	(.652)	.826
PREDICTED AREA QPA'S						
APQPA	115	(-)	.860	187	(-)	.823
BPQPA	51	(-)	.933	31	(-)	.922
HPQPA	149	(-)	.866	161	(-)	.884
HPQPA	37	(-)	.916	24	(-)	.862
AREA COMPATIBILITY STANDARD SCORES						
SIMILARITY OF GENERAL INTERESTS						
AGISS	639	(.5274)	.685	688	(.4702)	.614
BGISS	632	(.3744)	.646	656	(.1298)	.282
NGISS	639	(.5256)	.528	648	(.4675)	.592
HGISS	706	(.3038)	.494	755	(.1832)	.331
SIMILARITY OF OCCUPATIONAL INTERESTS						
AOISS	563	(.4983)	.647	580	(.5633)	.722
BOISS	582	(.3182)	.549	588	(.1815)	.395-
NOISS	565	(.5050)	.643	587	(.5552)	.703
HOISS	604	(.2708)	.440	618	(.1534)	.279
SIMILARITY OF ROKEACH VALUES						
ARVSS	589	(.2505)	.325	559	(.0377)	.048
BRVSS	589	(.2581)	.445	559	(.0358)	.079
NRVSS	589	(.2177)	.277	559	(.1040)	.132
HRVSS	589	(.1718)	.279	559	(-.0043)	.008

The first function, which clearly is Biglan's Established Paradigm dimension, dominates, accounting for 88% of the explained variance between groups on the components of the predicted QPA's. The second function conforms to his "Applied Science--Basic Science" dimension, and yields 10% of the explained variance. The USMA areas of concentration provide no groups for Biglan's "Living organism--nonliving" subject matter dimension.

The centroids on the first two discriminant functions for the four groups could conform to Guttman's 1954 criteria for a circumplex, but with only four points one cannot tell.

The AY 79-80 Study

a. Combining the Classes of 1978 and 1979 made possible "sharpening-up" the four Area Interest Profile Similarity Scales (IPSS) as well as the development of "Field of Study" IPSS equations. The mean IPSS scores made by the "high-half" and "low-half" groups presented the same general picture as that shown in Table 1 for area interest based scales. In developing these scales, a higher minimal level of significance was possible, and earlier "General Interest SCII scales were combined with the Occupation Interest scales to produce only one score per group. The result was a computer-produced profile for each cadet in the Class of 1982 which was distributed and used by their Academic counselors when providing guidance on the choice of field of study. (For a more detailed technical report, see Bridges, 1981. For less technical documentation, see Sliuford, 1980.)

b. More definitive information on the basic dimensions of the USMA curriculum resulted from the two equations for an Applied Science Raw Score (APSIRSC) and a Non-Physical Science Raw Score (NPSCIRSC), based solely on cadets' discriminating interest patterns. Correlations were computed between cadets' scores on these two scales and their IPSS scores for their own graduating area and field. The two were also correlated with the dummy variable for each of 36 area/field groups. The results are depicted in Figure 1. These data and an ISS intercorrelation matrix substantiates conformance to Guttman's circumplex criteria and to Biglan's model.

Figure 1 indicates that much of the basic simple structure of USMA's Areas of Concentration and Fields of Study seem to be described by the extent to which each is related to both an "Applied Science--Pure Science" dimension and a "Non-Physical Science--Physical Science" dimension. The locations for area dummy variables (01 to 05) are largely as expected, and dummy variables for their fields tend to cluster around them.

Two other characteristics warrant special comment. First, when the point plotted for each pair of dummy variable correlations is connected to the corresponding point for that group's ISS, the ray-like pattern depicts sharply the general similarity of the patterns of interrelationships for the two sets of variables. Second, the circular structure of the ISS pattern depicted in Figure 1 is known as a circumplex. The intercorrelations confirm that in general the closer the points, the higher the correlations between the ISS scores. Also, in accordance with Gutman's criterion (Gutman, 1954) the correlations between interest similarity scale scores tend to be negative for ISS's plotted directly across the circle from each other.

The basic structure of hte field ISS's clearly shows close conformance to their area's classifications in the Biglan model, with the single exception of the Economics field. The figure both provides further indication that the ISS's are valid and suggests further exploration of the use of these two dimensions to provide field-choice guidance to the cadets without even one "interpretative anomaly." A "Map of Fields" (a graph on which is plotted the centroids for each field) could be provided cadets together with their own APSCI and NPSCI scores. When the cadet plots his/her point on the "map," the fields which were closest to his/her own interest pattern would be clearly indicated. This might also be supplemented by a simple individual report of the cadet's discriminant scores for each field. These suggestions need to be checked further as the use of higher-order factors for practical individual

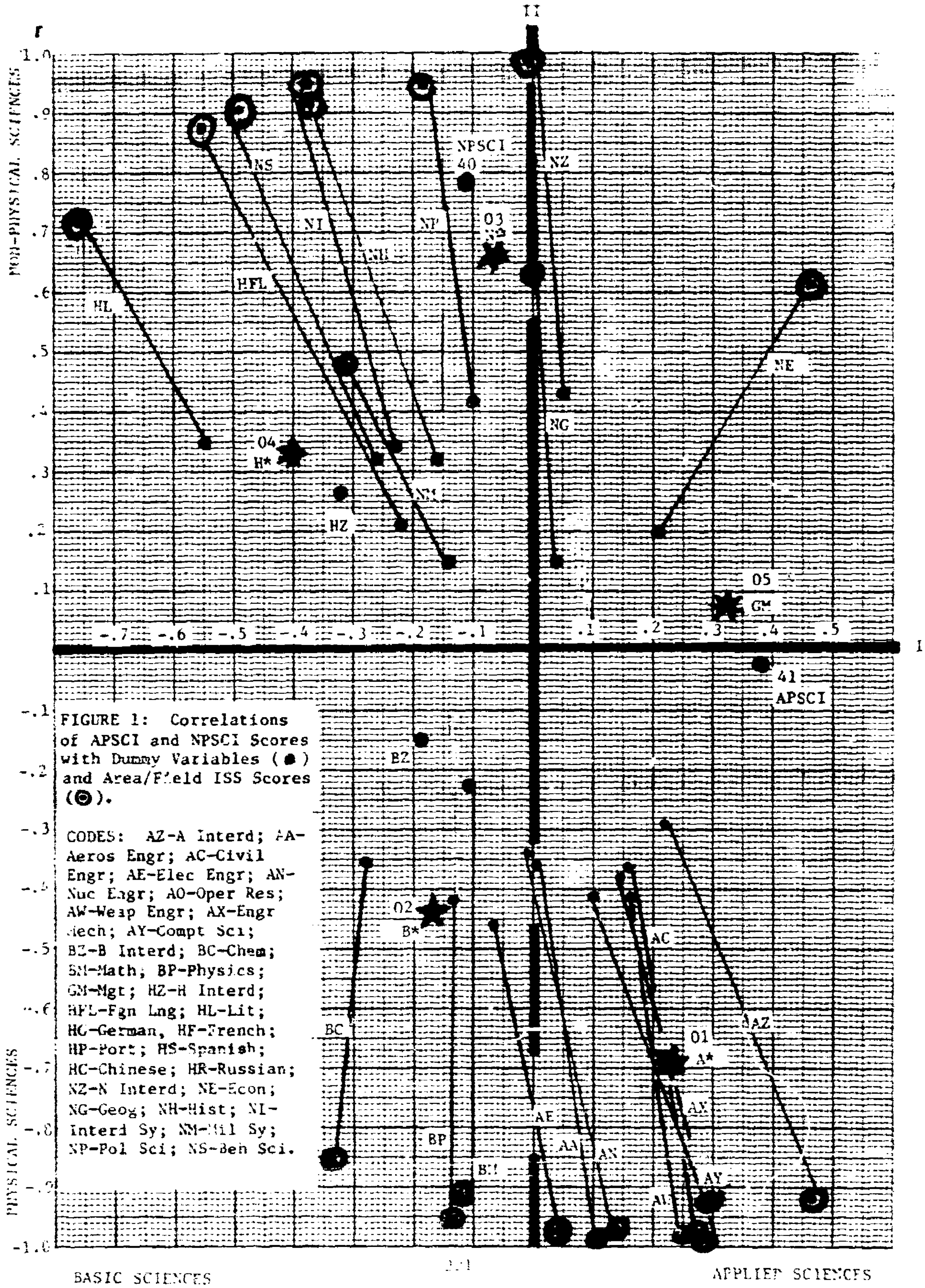


FIGURE 1: Correlations of APSCI and NPSCI Scores with Dummy Variables (●) and Area/Field ISS Scores (◎).

CODES: AZ-A Interd; AA-Aeros Engr; AC-Civil Engr; AE-Elec Engr; AN-Nuc Engr; AO-Oper Res; AW-Weap Engr; AX-Engr Mech; AY-Compt Sci; BZ-B Interd; BC-Chem; BM-Math; BP-Physics; GM-Mgt; HZ-H Interd; HFL-Fgn Lng; HL-Lit; HG-German, HF-French; HP-Port; HS-Spanish; HC-Chinese; HR-Russian; NZ-N Interd; NE-Econ; NG-Geog; NH-Hist; NI-Interd Sy; NM-Mil Sy; NP-Pol Sci; NS-Beh Sci.

prediction purposes often has been found to yield disappointing results--one reason for initially using it only as a backup procedure.

Incidentally, Army Service Branch IPSS scores developed from SCII scales for officers in the larger Service Branches also tend to form a circumplex. Factor analysis found two factors and oblique rotation indicated their negative correlation of -.52. The major dimension has Engineers, Ordnance, and Signal Corps at one end, and Transportation Corps and Quartermaster Corps at the other. It seems to be analogous to Biglan's established paradigm dimension. The second dimension has Military Intelligence at one end and Field Artillery at the other. This may be Biglan's "living organism--inanimate" dimension. It seems that the Army service branches studied do not have, as a whole, enough of the Applied-Pure differentiation for this dimension to show up in this factor analysis. None of the service branches stand out as "pure researchers."

The AY 80-81 Study

Addition of the cadets graduating in the Class of 1980 made possible the development of IPSS equations for additional fields of study, and the improvement of the eight field equations for which small interpretative anomalies had occurred previously. All interpretative anomalies have not been eliminated, but they have been reduced. High-half cadets in Physics and Computer Science still have a significantly higher mean IPSS for Mathematics than for their own field Cadets in Literature still have a higher mean on Foreign Languages IPSS than on Literature IPSS. Campbell (1977) stated that for some unknown reason, a few anomalous occupation criterion groups still were found to typically make a higher score on another occupation scale than on their scale. Perhaps further study of these field IPSS anomalies can throw more light on this situation.

The resulting computer-produced profile, sample shown as Figure 2, was distributed to the cadets in the Class of 1983 by the counselor involved in their choice of a field of study.

General Comments

The interest similarity profiles have been reported, by both cadets and academic counselors, as contributing to increased satisfaction and success of cadets in choice of area of concentration and field of study. The recent report of the Evaluation team representing the Commission on Higher Education of the Middle States Association of Colleges and Schools stated as follows:

"We think the academic advising process is outstanding. . . . The recent development which permits the Company Academic Counselor to use information developed through the Strong-Campbell inventory instrument is to be commended. This development, which requires far less training than the use of the inventory itself, enables the cadet to know how his or her interests correlate with the interests of those who have been successful in the various concentrations. This is a very helpful advising tool."

Conclusion

The scales developed by this approach provide valid and useful measures of the similarity of an individual to a criterion group on important group discriminating characteristics.

ISS PROFILE: AREA/FIELD INTEREST SIMILARITY SCALE SCORES

**CADET [redacted]
SSN [redacted] WHEN SELECTING YOUR FIELD OF STUDY THE INTEREST-BASED SCORES REPORTED BELOW
COMPANY [redacted] SHOULD BE CONSIDERED, AS SHOULD YOUR PAST SUCCESS IN FIELD-RELATED COURSES,
OTHER EVIDENCE OF YOUR RELEVANT SPECIAL ABILITIES, AND YOUR PERSONAL GOALS.

THESE INTEREST-BASED SCORES WERE DERIVED FROM YOUR RESPONSES TO THE STRONG-CAMPBELL INTEREST INVENTORY (SCII). THEY MAY BE INTERPRETED IN A MANNER SIMILAR TO THE SCORES ON YOUR SCII PROFILE. THESE SCORES, HOWEVER, INDICATE THE SIMILARITY BETWEEN YOUR INTEREST PATTERNS AND THE INTEREST PATTERNS OF CADETS GRADUATED IN THE UPPER HALF OF THEIR AREA OF CONCENTRATION OR FIELD OF STUDY. THE AVERAGE ISS SCORE FOR CADETS WHO DID WELL IN A PARTICULAR AREA OR FIELD IS 50.

IF YOUR SCORE FOR A PARTICULAR SCALE IS 65 OR HIGHER, THEN YOUR INTEREST PATTERN IS VERY SIMILAR TO THOSE OF CADETS WHO WERE SUCCESSFUL IN THAT AREA OR FIELD. THE HIGHER YOUR SCORE, THE MORE SIMILAR YOUR INTERESTS ARE TO THOSE OF SUCCESSFUL CADETS. THE LOWER YOUR SCORE THE LESS SIMILAR YOUR INTERESTS. IN FACT VERY LOW SCORES INDICATE DISSIMILARITY. **** IF YOUR SCORES FOR TWO OR MORE AREAS/FIELDS ARE FAIRLY CLOSE TOGETHER (SAY WITHIN 4 OR 5 POINTS), IT MEANS THAT YOUR INTEREST PATTERNS ARE ABOUT THE SAME FOR THOSE AREAS OR FIELDS. WHEN THE DIFFERENCE IS 5 OR MORE POINTS, YOU CAN HAVE CONSIDERABLE CONFIDENCE THAT YOUR INTEREST PATTERN IS MORE LIKE THAT OF CADETS IN THE AREA OR FIELD HAVING THE HIGHER SCORE. **** IF YOU SCORE HIGH IN AN AREA AND IN FIELDS WITHIN THAT AREA, YOU CAN HAVE MORE CONFIDENCE THAT THOSE AREAS/FIELDS WILL MATCH YOUR INTERESTS. PAY PARTICULAR ATTENTION TO THE FIELD WITH YOUR SIGNIFICANTLY HIGHER ISS(S) AND TO ANY FIELDS WITH SCORES ABOVE 65.

Scale: 20 30 40 45 50 60 70 80 — .....)(VERY.LOW)(...LOW..)(.AVERAGE)(..HIGH..)(VERY.HI.)(......

AREA/FIELD	CODE	INTEREST SIMILARITY SCORE	PROFILE
*APPLIED SCIENCES AND ENGINEERING	A*	(51)	A*
**MAS & E* INTERDISC FIELD	A	56	A
**AEROSPACE ENGINEERING	AA	49	AA
**CIVIL ENGINEERING	AC	56	AC
**COMPUTER SCIENCE	AY	47	AY
**ELECTRICAL ENGINEERING	AE	43	AE
**ENGINEERING MECHANICS	AK	52	AK
**NUCLEAR ENGINEERING	AN	50	AN
**OPERATIONS RESEARCH	AO	45	AO
**WEAPON SYSTEMS(MECHANICAL) ENGR	AW	47	AW
*BASIC SCIENCES	B*	(42)	B*
**"BS" INTERDISCIPLINARY	B	56	B
**CHEMISTRY	BC	40	BC
**MATHEMATICS	BM	47	BM
**PHYSICS	BP	32	BP
*HUMANITIES	H*	(26)	H*
**"H" INTERDISCIPLINARY	H	15	H
**FOREIGN LANGUAGES (SGL OR DBL)	FL	32	FL
**LITERATURE	HL	17	HL
*NATIONAL SECURITY-PUBLIC AFFAIRS	N*	(31)	N*
**INTERDISCIPLINARY	N	33	N
**BEHAVIORAL SCIENCE	NS	29	NS
**ECONOMICS	NE	37	NE
**GEOGRAPHY	NG	36	NG
**INTERNATIONAL AFFAIRS	NI	22	NI
**MILITARY STUDIES	NR	24	NR
**MODERN HISTORY	NH	19	NH
**POLITICAL SCIENCE	NP	20	NP
*MANAGEMENT(INTERDISCIPLINARY)	GM	(44)	GM

.....)(VERY.LOW)(...LOW..)(.AVERAGE)(..HIGH)(VERY.HI.)(......

FIGURE 2

REFERENCES

Biglan, A. "The Characteristics of Subject Matter in Different Academic Areas." Journal of Applied Psychology, 1973, 57, 195-203.

Bridges, C. F. The Value Profiles of Cadets. OIR Report #73-014, West Point, 1973.

Bridges, C. F. The Development of Guidance Materials for Cadets in the Selection of their Area of Concentration. OIR Report #79-006, West Point, 1979.

Bridges, C. F. Counselor's Evaluation of the USMA Academic Counseling Program. OIR Report, West Point, September 1980.

Bridges, C. F. The Development of Field of Study Interest Similarity Score Profiles. OIR Research Report #80-011, West Point, 1981.

Campbell, D. P. Manual for the Strong-Campbell Interest Inventory. Stanford: Stanford U. Press, 1977.

Claudy, J. G. "Biserial Weights: A New Approach to Test Item Option Weighting." Applied Psychological Measurement, 1978, 2, 25-30.

Creswell, J. W., and Bean, J. P. "Research Output, Socialization, and the Biglan Model." Research in Higher Education, 1981, 15, 69-91.

Gordon, Leonard. The Assessment of Interpersonal Values. Chicago: Science Research Associates, 1975, p. 84.

Guttman, L. New approach to factor analysis: The radex. In P. F. Lazarsfeld (Ed.), Mathematical Thinking in the Social Sciences. Glencoe, IL: Free Press, 1954.

Holland, J. L. Manual for the Self-directed Search. Palo Alto, California, Consulting Psychologists Press, 1980.

Rokeach, Milton. The Nature of Human Values. New York: MacMillan, The Free Press, 1973.

Shuford, MAJ J. H. The Development of an Interest Based Guidance Profile for Use in Academic Counseling. OIR Report #81-002, West Point, 1980.

AD P001297

Brittain, Clay V. & Brittain, Mary M., Williamsburg, Virginia. (Fri. A.M.)

Reading Ability, Readability, Motivation and Test Validity

According to a prevalent argument about occupational proficiency testing, multiple-choice tests have built-in sources of invalidity. One such source pertains to the reading competencies required in taking a multiple-choice test. The test, it is argued, requires reading competencies beyond those required by the job itself. Thus, one finds job incumbents whose job performance is outstanding but who cannot score well on a paper-and-pencil test. There is also a rather commonly accepted argument about the interactive effects of reading ability and motivation on performance. How well one handles written material is a function of motivation as well as reading ability. The present paper will analyze the correlation of test performance and reading ability in relation to the readability of the test materials and the motivation of examinees.

Reading Ability, Readability, Motivation and Test Validity

Clay V. Brittain
Army Training Support Center

Mary M. Brittain
Virginia Commonwealth University

This study was concerned specifically with the performance of soldiers on Skill Qualification Tests (SQT) in relation to the readability of the tests. More generally, the paper also touches upon the effects of soldiers' motivation and training on SQT performance.

First, we will comment briefly about the makeup and scoring of SQT. Skill Qualification Tests are made up of the following parts, or components: the Job Site Component (or JSC), the Hands-On Component (or HOC), and the Skill Component (or SC).

The Job Site Component (JSC) is not a test in the conventional sense. Rather, supervisors rate soldiers on their ability to perform certain tasks. The Hands-On Component (HOC), as the name implies, is made up of performance tests. Soldiers perform certain tasks under a set of standard conditions and their performance is scored according to clearly specified criteria. The Skill Component (SC), as the name does not imply, is a multiple-choice test. The SQT score is an aggregate of the component scores. The written portion of the SQT (the Skill Component) is the major source of variance in SQT scores; and, of course, this is the component where the problem of readability arises.

One of the major concerns about printed multiple-choice tests of occupational competence is that the test may require reading ability in excess of that required by the job itself. If an SQT requires reading skills which are excessive in relation to the reading abilities of soldiers and the demands of the MOS, then there obviously are serious implications for test validity.

Soldiers with limited reading skills may not be able to achieve test scores which are commensurate with their MOS competence.

Major Questions and Methodology

With respect to readability, the study was designed to address three questions as follows:

(1) Does the readability of SQT match the readability in the target soldier population?

(2) To what extent is reading ability a factor in the SQT performance of soldiers?

(3) To what extent is the association between reading ability and SQT performance modified by other factors, specifically soldier motivation?

These questions were addressed here through analysis of ten SQT and data relating thereto. The SQT were all skill level one tests. Skill level one is the lowest of the five skill levels in the Army MOS structure. The SQT were drawn from the ten MOS shown in Table 1. The first five MOS listed here are Combat Arms MOS. The latter five are Support MOS (i.e., Combat Support/Combat Service Support). This mix of MOS was by design.

(Table 1 About Here)

As part of a larger study, data for use in evaluating these ten SQT were collected from skill level one soldiers who had taken the SQT, and from their supervisors. The SQT also were reviewed by specialists from the civilian testing community. The paper draws upon data provided by the field research and the expert reviews of the SQT.

Match Between SQT Readability and Soldier Reading Ability

The readability of the SQT, expressed as a reading grade level (RGL), was estimated through application of the Flesch-Kincaid reading ease formula which uses sentence length and word length to estimate the readability of prose materials (Kincaid & Fishburne, 1977).

The reading ability of soldiers, also expressed as a reading grade level (RGL), was estimated from AFQT scores. The average reading ability in each MOS population was estimated using a conversion scale based upon correlations of AFQT scores with scores on the Nelson-Denny Reading Test and the Gates-MacGinitie Reading Test. This conversion scale was developed by the Air Force Human Resources Laboratory (Mathews, Valentine & Sellman, 1978). Estimates of the RGL of each SQT and the RGL of the soldier population for which the SQT was intended are shown in Table 2. These estimates indicate consistent matches between SQT readability and soldier reading ability. The average RGL of the SQT generally is below the average RGL of the soldier population. The only exception to this match of SQT/RGL and soldier/RGL is the SQT for MOS 55B.

(Table 2 About Here)

The conclusion indicated here is that readability is not a major source of invalidity in the SQT. The conclusion also is consistent with data from the interviews of noncommissioned officers.

When asked for their opinions as to the major reasons why soldiers fail SQT, NCO's only infrequently mentioned reading ability.

Reading Ability as a Factor in SQT Performance

It should not be concluded, however, that reading ability is unimportant in SQT performance. A questionnaire which was completed by soldiers included the item "I have difficulty understanding the written portion of my SQT." Soldiers responded by checking one of the following: never, seldom, sometimes, usually, or always. The percentage of soldiers in each MOS who responded with "sometimes," "usually" or "always" is shown in Table 3. There is a substantial proportion of soldiers who reported difficulty in understanding the skill component

(Table 3 About Here)

Soldier responses to the item "I have difficulty understanding the written portion of my SQT" were quantified and correlated with SQT scores. The correlations, while not all statistically significant, were all negative. Finally, we computed correlations between AFQT scores and SQT scores and regarded these correlations as estimates of the contribution of reading ability to SQT variance. The correlations are shown in Table 4.

(Table 4 About Here)

It is pertinent to note here that on each SQT, the reviewers found items which were complex and unnecessarily difficult in wording. In a number of instances, these items were answered correctly by even fewer soldiers than would be expected on the basis of chance.

The conclusion, stated above, that readability of the test is not a serious problem must then be qualified. Readability of the test and reading ability of soldiers clearly are factors in SQT performance.

Effects of Soldier Motivation

We now wish to address the third question: to what extent is the association between reading ability and SQT performance modified by other factors, especially soldier motivation?

(Figure 1 About Here)

The question stems from a simple model of reader performance (Klare, 1973), shown in Figure 1. It says, in effect, the reader's performance depends upon the interaction of three factors: the reader's level of competence, the readability of the materials, and the level of reader's motivation. Thus, a highly motivated soldier might effectively deal with materials at a level of readability that the less well motivated soldier of comparable reading ability could not deal with effectively. This was an area which we wished to explore.

As a means of assessing motivation, soldiers responded to the following questions:

How important is to you to get a good score on your SQT?

How important do you feel the SQT is for promotion?

Do you intend to make the Army a career?

Would you recommend your MOS to someone considering a career in the Army?

It was expected that soldiers who expressed their intentions to make a career of the Army and those who would recommend their MOS to someone else would probably be more strongly motivated to do well on the SQT. It was also assumed that those who said that the SQT was important tried harder to do well. But when SQT scores were analyzed in relation to soldiers' answers to these questions, no relationship wa found. The absence of an association between SQT scores and soldier motivation as measured here was unexpected. It may be that our questions simply did not reflect soldier motivation. But another explanation, and in our opinion, a more plausible one, has to do with the experience of taking a test. For most people in our society, taking a test is a stressful situation in which one tries to do their best. We suspect that this also is true of soldiers taking an SQT. Thus, differences in motivation, such as our questions were designed to probe, are overwhelmed and washed out in the test situation itself.

This suggests that differences in motivation impact on test scores, if at all, through prior efforts to prepare for the test. But among skill level one soldiers, training for the SQT seems not as much a matter of individual motivation as of unit emphasis. Two indices of unit training are the timely delivery of SQT Notices to soldiers and the amount of training for the SQT. Both of these indices are positively correlated with the SQT scores as shown in Table 5.

(Table 5 About Here)

The SQT Notice informs soldiers on just what tasks are to be tested on the SQT. The correlations between training time and SQT scores may seem surprisingly low. But an important element here is the selectivity in training soldiers for the SQT. The training, at least in many instances, is based upon diagnostic information about soldier performance. Those soldiers who are identified as nonperformers of a task tend to get the most training.

Summary and Conclusions

In summary, the major results of this analysis are as follows:

1. The readability of the ten SQT included in this study generally match the reading ability in the target soldier population. Thus, readability would not seem to be a serious source of invalidity in SQT.

2. However, soldier reading ability and SQT readability are important factors in SQT performance. Evidence for this conclusion comes from reviews of SQT by outside experts, from soldier responses indicating difficulty in understanding the skill component of the SQT, and correlations of AFQT and SQT scores.

3. Individual soldier motivation as expressed in military career intentions, attitudes toward the MOS, and opinions about the importance of the SQT was not shown to be related to SQT scores.

4. Unit emphasis on the SQT as reflected in the early delivery of the SQT Notice and training of soldiers was related to SQT scores.

5. With respect to the question of major interest here, the findings are negative. This question concerned the association between soldier reading ability and SQT performance, as the association might be modified by other factors. There is evidence that SQT scores are associated with reading ability. But our data did not give evidence that this association is modified by soldier motivation or unit emphasis on the SQT.

References

Kinkaid, J. P., & Fishburne, R. F. Reading Formulas for Military Training Materials. Human Factors Society Bulletin, 1977, vol. 20, no. 7.

Klane, G. A Manual for Readable Writing. Glen Burnie, M.D. REM Co. 1975.

Mathews, J. J., Valentine, L. E. Jr., & Selman, W. S. Prediction of Reading Grade Levels of Service Applicants from Armed Services Vocational Aptitude Battery (ASVAB). Air Force Human Relations Laboratory, Technical Report No. 78-82. December 1978.

TABLE 1

SQT SELECTED FOR ANALYSIS

SQT	JOB TITLE	PROPONENT SCHOOL
11H	HV ANTIARMOR WPN CRWMN	INF
12C	BRIDGE CRWMN	EN
13B	CANNON CRWMN	FA
16P	SHT RGE MSL CRWMN	AD
19E	M48/M60 ARM CRWMN	AR
31M	MULTI-CHAN COMM EQ OP	SIG
55B	AMMO SP	MMCS
67N	UTIL HEL REP	TRANS
71L	ADMIN SP	SSC
76X	SUBSIST SUP SP	QM

Table 2

MATCH OF SQT READABILITY WITH

SOLDIER READING ABILITY

MOS	SQT READING GRADE LEVELS	SOLDIER READING GRADE LEVELS
11H	6	9
12C	5	8
13B	6	8
16P	8	8
19E	6	8
31M	7	8
55B	9	8
67N	8	10
71L	8	11
76X	9	8

Table 3

CORRELATION OF REPORTED DIFFICULTY WITH SOLDIER SCORES ON THE WRITTEN PORTION OF THE SQT

MOS	CORRELATION WITH SC SCORES*
11B	-.30
12C	-.34
13B	-.39
16P	-.04
19E	-.09
31M	-.02
55B	-.18
67N	-.19
71L	0.00

*Negative correlations indicate that some difficulty resulted in lower scores.

TABLE 4

CORRELATIONS OF SQT SCORES WITH AFQT SCORES

MOS	N	SQT SCORES	SC SCORES
11H	31	.43*	.41*
12C	12	.26	.47
13B	27	.42*	.39*
16P	22	.22	.32
19E	55	.22	.37*
31M	17	.48*	.41*
55B	27	.30	.25
67N	46	.22	.21
71L	27	.31	.35*

*STATISTICALLY SIGNIFICANT

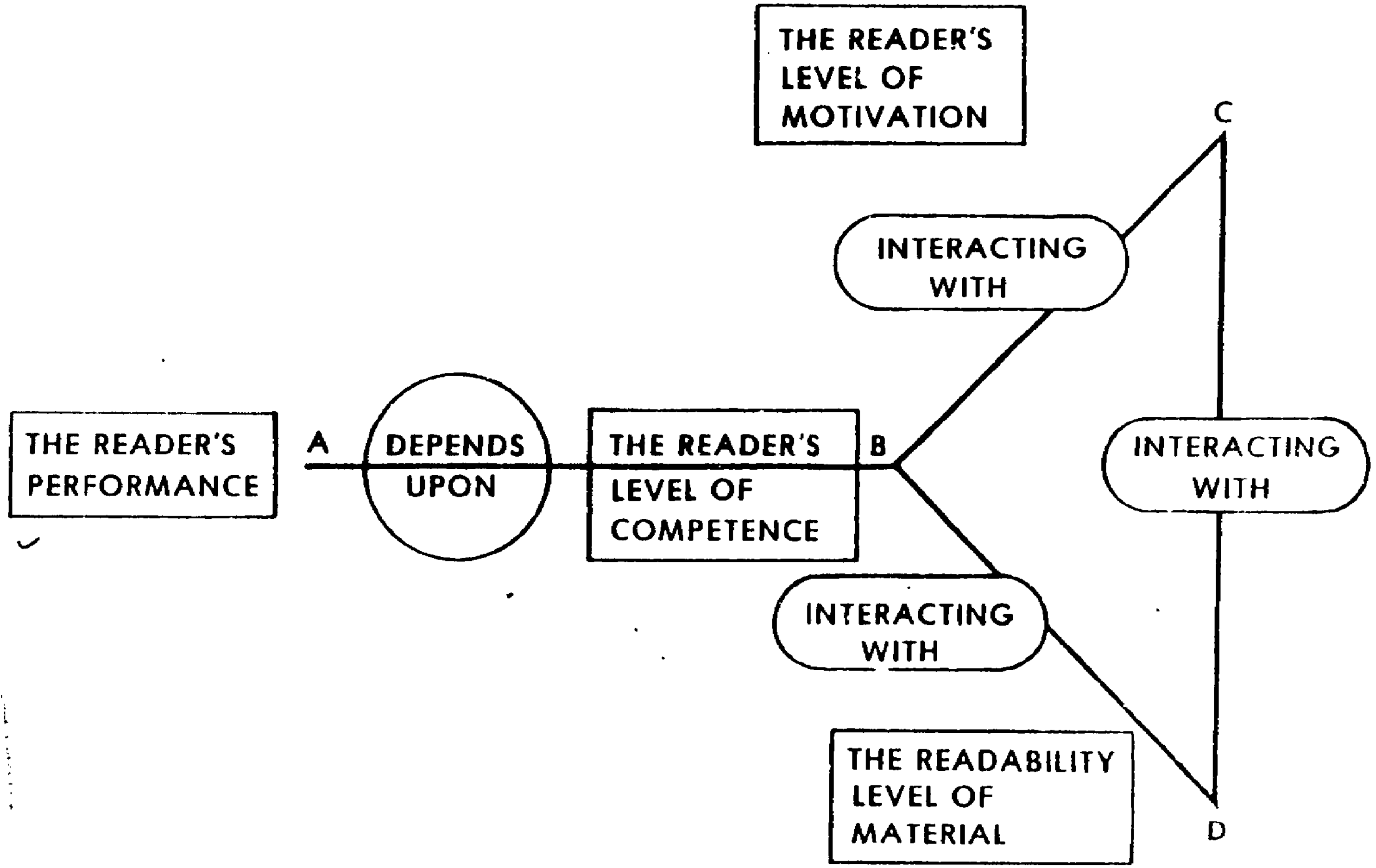
Table 5

CORRELATION OF SQT SCORES WITH NUMBER OF WEEKS SOLDIERS HAD SQT NOTICE AND NUMBER OF HOURS SPENT TRAINING

MOS	SQT NOTICE	TRAINING TIME
11H	.43*	.40*
12C	.61*	.21
13B	.24	.28*
16P	-.06	.25*
19E	.28*	.09
31M	.15	.13
55B	.40*	.22
67N	.15	-.02
71L	.14	.26
76X	Insufficient Data	.21

*Significant at .05 level

FIGURE 1


THE READER'S PERFORMANCE
A
DEPENDS UPON
THE READER'S LEVEL OF COMPETENCE
B
INTERACTING WITH
THE READER'S LEVEL OF MOTIVATION
C
INTERACTING WITH
INTERACTING WITH
THE READABILITY LEVEL OF MATERIAL
D

AD P001298

FORECASTING ARMY OFFICER RETENTION PRIOR TO COMMISSIONING[1]

Richard P. Butler, Ph.D., Chief, Research Branch
Mr. Claude F. Bridges, Research Psychologist
Mr. John W. Houston, Statistician

Office of the Director of Institutional Research
United States Military Academy
West Point, New York 10996

Valid selection instruments may help the Army to diminish the problem of retaining qualified officer personnel after their initial tours of duty. This report describes, validates, and updates information on the Military Career Commitment Gradient (MCCOG), an instrument designed to predict, prior to commissioning, whether an individual will remain on active duty after his initial tour. The MCCOG was administered to the U.S. Military Academy Classes of 1966, 1967, and 1969 shortly before commissioning, and to the Classes of 1970, 1971, and 1972 one, two, and three years, respectively, before commissioning. The MCCOG scores for all six classes were significantly correlated with the criterion--on active duty versus not on active duty--gathered seven to nine years later. The coefficients ranged from .17 when the MCCOG was given three years prior to commissioning, to .54 when the MCCOG was administered just prior to commissioning. Possible uses of the MCCOG for any commissioning program are discussed.

[1] Any conclusions in this report are not to be construed as official U.S. Military Academy or Department of the Army positions unless so designated by other authorized documents.

FORECASTING ARMY OFFICER RETENTION PRIOR TO COMMISSIONING

A professional officer corps can be developed and maintained only by retaining those officers who are qualified and motivated to lead. Perhaps no other aspect of developing a professional officer corps has been of greater concern than has the problem of retaining qualified personnel after their initial tour of service (Edmonds, 1972). One method of overcoming this problem is to identify, prior to commissioning, those individuals who are not likely to remain on active duty after their initial tour of service. Preventing commissioning or career counseling and programs designed to increase motivation for a service career could then be started, if thought useful.

Previous research to identify, prior to commissioning, individuals likely to leave active duty upon completing their initial tour has not been prevalent. The only well-controlled studies that the authors are aware of were completed by Shenk (1972, 1973) and Butler and Bridges (1978). Shenk investigated the officer input from the various Air Force commissioning programs concerning the predictability of their future status (on active duty versus not on active duty). The major finding was that the correlation between a five-point career intent question and status five years later was .24, a coefficient that Shenk considered rather low. Butler and Bridges, using a one-question instrument called the Military Career Commitment Gradient (MCCOG) for two classes at the U.S. Military Academy, found that it had significant correlations of .54 and .39 with status seven years later.

The purpose of this report is to describe, revalidate, and update information on the MCCOG, an instrument designed to predict, even before commissioning, the probability of an individual remaining on active duty beyond his/her initial tour.

METHOD

Instrument

The MCCOG (Table 1) is a one-question instrument designed to measure precisely the strength of an individual's commitment to a military career. The rationale upon which various characteristics of the MCCOG are based is organized around the major problems of measurement peculiar to rating scales--such as controlling the set, providing meaningful and practical scaling of scores, and insuring that respondents rate the same thing and perceive and use the same scale. Of special import is the length of the response scale. Instead of the usual two to five defined areas as alternatives, the scale consists of 99 scaled points on a gradient with definitions placed at 19 locations in terms of the normal probability distribution (like T-score percentiles), and with 13 verbal descriptions selected on the basis of empirical data as to relative level of meaning consistently implied.

With such a large number of continuous scaled response points, reliability is quite high. The reliability coefficient of the MCCOG scores, based on test-retest data, freed of temporal change effects (Heise, 1969), was .83 for a sample of 564 respondents. Comparisons of validity data on MCCOG, on the usual five alternatives to essentially the same question, and on a semantic differential type scale developed by the ARI Sample Survey Group, suggested that the much higher validity of the MCCOG is at least partially due to its high reliability.

Table 1. Military Career Commitment Gradient

This item concerns the intensity of your desire for a career as an officer in the military service. It consists of (1) a question and (2) a response gradient extending continuously between two defined extreme values.

Selected areas on the gradient are described, both verbally and in terms of probabilities, to provide you with some meaningful reference points and to provide for more precision in scalar interpretation.

At selected scalar points, percentages beside the gradient indicate the judged probability (number of judged chances in 100) of one voluntarily continuing his active military career until mandatory retirement. Note, however, you definitely should NOT limit yourself to the few points for which descriptions are provided.

Because of the procedures for analyzing this item, it is very important that you follow these instructions precisely, step by step.

INSTRUCTIONS. Complete each step before going to the next one.

Step One. Thoughtfully read the question in the box below:

QUESTION: To what degree are you now certain that you will continue an active military career until mandatory retirement?

Step Two. At the bottom of the gradient, on the opposite page, read the defintion of that extreme point on the gradient.

Step Three. At the top of the gradient, read the definition of that extreme point.

Step Four. At the middle of the gradient, the 50% probability point, read the description of that point.

Step Five. Locate the general area on the gradient which seems to correspond best with your current commitment; thoughtfully read the descriptions of the near points and check the space on the gradient that most closely represents your current level of commitment. Do NOT limit yourself to the few points described verbally.

Step Six. Select the coded letter and number combination at the left of the checked space on the gradient. Enter this as illustrated below. For example, if you had checked the space coded "f1," you would mark the answer sheet as follows:

(Continuing Table 1)

CODE	MILITARY CAREER COMMITMENT GRADIENT
j5	"+ ∞"–There is infinite probability that I will continue my active military career as long as I possibly can, a career as an officer in active military service is more important to me than is anything else in the world. There is absolutely no chance at all that anything in the world could ever develop that could cause me to voluntarily resign.
j4	
j3	
j2	
j1	
i5	99.995%
i4	
i3	
i2	
i1	99.9%–I am virtually certain that I will continue my active military career as long as I am allowed to do so--that I will NOT voluntarily resign.
h5	
h4	
h3	
h2	99%—I am almost certain I will make a continuing military career if possible.
h1	
g5	
g4	95%
g3	
g2	90%—I am confident that I will make a continuing military career and NOT voluntarily resign.
g1	85%
f5	
f4	75%—I am very likely to continue my military career as long as possible.
f3	
f2	65%—I probably will remain in the military service after completion of my military obligation as an officer.
f1	
e5	50%—I am not inclined the least bit either way at present.
e4	
e3	35%—I am not sure but probably will resign after completing my military obligation as an officer.
e2	25%—I am very likely to resign when I can honorably do so after completing my military obligation as an officer.
e1	
d5	15%
d4	10%—At this time, I am confident I will resign my commission after completing my military obligation.
d3	
d2	5%
d1	
c5	
c4	1%—As of now, I am almost certain that I will get out of the military service as soon as I possibly can.
c3	
c2	
c1	
b5	0.1%—I am virtually certain that I will resign when I can.
b4	
b3	
b2	
b1	0.005%
a5	
a4	
a3	
a2	
a1	"– ∞" In my personal feelings, attitudes and thoughts, I am utterly committed to a completely non-military occupational career and life as soon as it is at all possible. There is absolutely no possibility whatsoever that I will continue as an officer in the military service beyond my minimal obligated military duty.

The MCCOG was developed to reflect the result of the interaction of both intent to pursue and intensity of desire for a military career. Focusing the instrument on expressed intentions is supported by a literature review concerning employee turnover by Porter & Steers (1973). They found that expressed intentions concerning future participation in an organization appeared to be a better predictor than measures of job satisfaction. For example, Porter & Steers cited Kraut (1970) who, in a study of managerial personnel, found significant relationships between expressed intent to stay and subsequent employee participation. The correlations were far stronger than those found between expressed satisfaction and continued participation. Furthermore, Atchison & Lefferts (1972), in a study of turnover among Air Force pilots, found that turnover was significantly related to the frequency with which pilots thought about leaving their jobs.

Subjects and Procedure

The MCCOG was administered shortly before commissioning, which takes place at graduation, to cadets in the U.S. Military Academy Classes of 1966, 1967 and 1969, and in August 1969 to the Classes of 1970-71-72. The advantage of the August 1969 administration was that it allowed the predictive power of the MCCOG to be assessed for Classes which were one, two, and three years from graduating and commissioning. The results of the Classes of 1966 and 1967 administrations were reported earlier by Butler and Bridges (1978), but will be included in the current report to give a more complete picture.

Usable returns that could be paired with the criterion of status were received from 396 graduates from the Class of 1966, 465 (Class of 1967), 550 (Class of 1969), 539 (Class of 1970), 611 (Class of 1971), and 708 (Class of 1972). The samples were less than the number of graduates in each class because only those graduates with MCCOG scores who were commissioned in the Army and were still on active duty, or who were commissioned in the Army and had voluntarily (unqualified) resigned from active duty were included in the analysis. Those graduates who were allied cadets, commissioned in other services, involuntarily separated, retired for disability, or otherwise separated (including deaths), were not included in the samples. Graduates from the Classes of 1966 and 1967 incurred mandatory four-year military obligations (the other classes had five-year obligations), after which they were free to resign their commissions, unless they had incurred additional obligations because of additional education.

The status criterion for the Classes of 1966 and 1967 was obtained in 1973 and 1974, respectively, seven years after the MCCOG was given. The criterion for the remaining classes was gathered in 1978, nine years after the MCCOG was administered. Thus, the time after completion of their mandatory obligation ranged from one year for the Class of 1972 to four years for the Class of 1969.

Data Analysis

Point-biserial correlations and expectancy tables were used to express the relationships between MCCOG scores and status (scored 0/1). To correct the correlations to a common base with a possible maximum correlation of 1.00, the obtained values were divided by the maximum r_{pb} possible for the proportions in the two groups.

RESULTS

Table 2 shows that the relationship between MCCOG scores and status for all class years is statistically significant. As one might expect, the correlations were higher for Classes (1966-67-69) administered the MCCOG just prior to graduation/commissioning.

Table 2: Relationship Between MCCOG Scores and Status

Class Year	Status	MCCOG N	MCCOG $\bar{X}$	MCCOG SD	Corrected r_{pb}
1966	Active	257	60	11	.54***
	Resigned	139	45	12	
1967	Active	271	59	11	.39***
	Resigned	194	51	13	
1969	Active	279	56	12	.35***
	Resigned	271	49	13	
1970	Active	323	50	16	.26**
	Resigned	216	43	17	
1971	Active	321	54	16	.23**
	Resigned	290	48	18	
1972	Active	409	56	16	.17*
	Resigned	299	52	15	

*p < .10; **p < .02; ***p < .001

The expectancy tables shown in Tables 3, 4, and 5 give clear illustrations of the relationships between MCCOG scores and status. They indicate that individuals in the Classes of 1966 and 1969 who scored 61 or above on the MCCOG

Table 3. Relationship Between MCCOG Scores and Status
(Class of 1966)

MCCOG Score	Percent Remaining on Active Duty (0–100)	Number Scoring in this Range
81 or over	100%	13
71-80	93%	30
61-70	83%	84
51-60	67%	159
41-50	42%	93
1-40	6%	17

Table 4. Relationship Between MCCOG Scores and Status
(Class of 1969)

MCCOG Score	Percent Remaining on Active Duty	Number Scoring in this Range
81 or over	64%	14
71-80	74%	31
61-70	75%	57
51-60	56%	263
41-50	35%	130
1-40	22%	55

Table 5. Relationship Between MCCOG Scores and Status
(Class of 1972)

MCCOG Score	Percent Remaining on Active Duty	Number Scoring in this Range
81 or over	80%	59
71-80	64%	45
61-70	60%	89
51-60	58%	228
41-50	52%	210
1-40	49%	77

were very likely to be still on active duty, and those scoring 50 or below were likely to have resigned from active duty. For the Class of 1972 the relationship is not as sharp, but there is a definite tendency for individuals to remain on active duty as MCCOG scores increase. For brevity, the data from the Classes of 1967, 1970, and 1971 are not presented, but the results were similar to the classes that were presented.

DISCUSSION

The prime purpose of this study was to determine the validity of the MCCOG, when given prior to commissioning, as a predictor of retention on active duty beyond the obligatory first tour of duty. The results indicated that the MCCOG is a valid predictor of retention, even when administered several years prior to commissioning and even when the criterion is gathered nine years later.

The findings support the theory that the individual's direct estimate of his future tenure is a good predictor of turnover, especially when determined with high precision. Kraut (1975) states that the individual himself is the best vehicle for properly weighting and integrating the factors that go into his decision to quit or remain in a job. He believes that the individual can provide the best synthesis of attitude toward his work situation, his opportunities elsewhere, and other aspects of his life that bear on his decision to remain in a job. The MCCOG follows Kraut's logic quite well. It is a direct measure of intentions and calls for the respondent to integrate all the various aspects that influence his/her intent to be a career Army officer.

The findings imply that the MCCOG might be a useful instrument for several purposes for any source of commissioning. First, in research for which the ultimate criterion is whether or not an individual will remain on active duty, the MCCOG can function as an intermediate criterion that is available prior to commissioning. Second, as a measure of commitment, it might be a valuable tool in evaluating programs designed to increase motivation for a service career. For example, the MCCOG might be administered before and after a summer training program. Third, in manpower planning, the MCCOG scores of a group give a fairly accurate appraisal of the number of individuals that are likely to remain on active duty. Fourth, the MCCOG might be an aid in the career counseling process of individuals. A fifth potential utilization is in the admissions area, but only if the MCCOG can be corrected for the expected high response bias, could it be administered to applicants. However, it could serve as the criterion to be predicted by another instrument that could help to differentiate those applicants who are likely to develop high commitment to a military career as opposed to those who are not.

REFERENCE NOTES

1. Edmonds, M. O. An examination of factors relating to retention of U.S. Military Academy graduates. Army War College, Carlisle Barracks, PA, 1972.

2. Shenk, F. Predictability of expressed career interest (Report No.: Tech. Rpt. 72-25). Air Force Human Resources Laboratory, Lackland Air Force Base, Texas, 1972.

3. Shenk, F. Development and validation of scores to predict officer career status (AD-760-540). Air Force Human Resources Laboratory, Lackland Air Force Base, Texas, 1973.

REFERENCES

Atchison, T. J., & Lefferts, E. A. The prediction of turnover using Herzberg's job satisfaction technique. Personnel Psychology, 1972, 25, 53-64.

Butler, R. P., & Bridges, C. F. Prediction of officer retention prior to commissioning. Journal of Occupational Psychology, 1978, 51, 177-182.

Heise, D. R. Separating reliability and stability in test-retest correlation. American Sociological Review, 1969, 34, 93-101.

Kraut, A. Predicting turnover of employees from measured job attitudes. Organizational Behavior and Human Performance, 1975, 13, 233-243.

Porter, L. W., & Steers, R. M. Organizational work, and personal factors, in employee turnover and absenteeism. Psychological Bulletin, 1973, 80, 151-176.

AD P001299

US ARMY TRADOC SYSTEMS ANALYSIS ACTIVITY
(USATRASANA)
CAVALRY FIGHTING VEHICLE (CFV)
FORCE DEVELOPMENT TEST AND EXPERIMENTATION (FDTE)
TRAINING ANALYSIS

BY
WALTER G. BUTLER
RONALD B. COOPER

1. SUMMARY.

This report presents the results of the training effectiveness analysis conducted on a training program designed to transition train an armored cavalry from the M113A1 to the CFV. Soldiers in the 19D (armored cavalry scout crewman) were trained in the operation of the CFV and its major systems in preparation for participation in an FDTE at Ft Knox. Data analyzed included results of written and hands-on tests, results from training on the prototype Conduct of Fire Trainer (COFT) and results from live fire exercises involving the 7.62mm coaxially mounted machine gun and the 25mm automatic cannon. Relationships between these data and soldier demographics/aptitude/attitude were investigated. Problem areas were observed in the soldiers' proficiency at acquiring, engaging and hitting targets with the CFV's weapons systems. Implications into CFV crewman soldier selection criteria were gained from this analysis.

2. BACKGROUND.

a. The CFV is designed to replace the M113A1 as the armored cavalry's fighting vehicle. It is basically the same as the Infantry Fighting Vehicle (IFV) with the rear compartment configured to accommodate an increased basic load of TOW missiles and a decreased number of passengers. (The IFV will carry seven men in the rear compartment; the CFV two). The CFV is armed with a 25mm automatic cannon, a coaxially mounted 7.62mm machine gun and a two-pod TOW launcher.

b. The IFV underwent an Operational Test (OT) II from 15 October 1979 to 26 November 1979. Cavalry operations were not examined in enough detail during this test and a CFV FDTE was scheduled at Ft Knox, KY from 2 June 1980 through 8 August 1980 to address these items. A cavalry platoon stationed at Ft Knox was selected to participate in the FDTE. This platoon was judged ARTEP-qualified on the M113A1 prior to beginning the training program designed to transition it from the M113A1 to the CFV. Five crews of five men each were organized from the platoon's 19D MOS (cavalry scout crewman) personnel to be trained as CFV crewmen. The five crew positions are: track commander (TC), gunner, senior scout (SS), junior scout (JS) and driver. These twenty-five crewmen are referred to in this report as "players".

c. The training period was divided into two phases, individual and collective training, with each phase lasting approximately three weeks. During individual training, all crewmen received the same instruction conducted primarily in a classroom or motor pool. Training during the collective phase was more specific to duty position and was conducted on tactical ranges.

3. SCOPE

a. The individual training was primarily a combination of lecture, demonstration and practice. Training devices and aids used were the CFV, the Conduct of Fire Trainer (COFT), printed handouts and draft manuals. At the conclusion of a training session or block of instruction, a hands-on test was given. The tests were administered by instructor personnel and were scored on a Go/No-Go basis. Personnel who failed to achieve a GO on a test were required to prepare themselves on their own for a retest. A second failure resulted in instructor assistance to prepare for another retest. At the end of the individual phase of training, the more important of these hands-on tests were given again in a two-day comprehensive examination.

b. Training on the COFT was conducted during the individual phase with all twenty-five players participating. Each player was presented a series of targets on the COFT and the results of his engagements were recorded. Not all players received the same amount of COFT training due to player absences and time constraints.

c. The collective training phase was conducted on tactical ranges at Ft Knox. During this phase of instruction the players participated in a series of exercises which required the performance of several tasks to meet the objective(s) of the exercise. For example, to successfully complete the range determination exercise, a crew was required to acquire a target within a given time standard and the gunner was then required to determine the range to the target within another set of standards.

d. These exercises were used as the hands-on tests during the collective phase. However, in most cases personnel who failed to achieve a "Go" on a given exercise were not re-tested. This was due to time constraints beyond the control of the instructors. During this phase, the crews also participated in live firings of the 7.62mm COAX, the 25mm main gun and the TOW missile.

4. INDIVIDUAL TRAINING PHASE.

a. At the end of a task training session, the students were tested individually on the task(s) covered. The tests, referred to as "Class HOT", were hands-on and were scored as "Go" or "No Go". At the end of the individual phase, the majority of these tasks were re-tested. This group of tests are referred to as "comprehensive HOT" or "Comp HOT".

b. Data from these tests show that there were 498 Go's on 507 first tries of the class hands-on tests which gives a first-time-Go rate of 98%. For the comp HOT, there were 349 Go's on 364 first tries for a first-time-Go rate of 96%. These rates indicate that the training provided the necessary skills to the players for successful performance of these tasks at the time they were tested. Problems were observed during subsequent training (especially the COAX and 25mm live fire exercises) involving the following tasks taught during this phase:

- Loading the COAX and 25mm weapons.

- Performing immediate action and misfire procedures on the COAX and 25mm weapons.

c. In loading the 25mm gun, the players experienced difficulty in properly connecting the ammo feed chutes. Improper connection of these chutes caused many stoppages which had to be corrected by the instructors or, in some instances, contractor representatives. The players were also unable to consistently perform the misfire and immediate action procedures in the proper sequence. More training and/or practice at the performance of these tasks are needed so that the skills can be retained by the soldiers. The connection of the ammo feed chutes should be especially addressed because of the manipulation skills required.

d. During the individual training phase the students were tested and determined to be representative of the overall 19D MOS Army population in terms of aptitude and demographics.

5. COLLECTIVE TRAINING PHASE.

a. During the collective phase of training, tasks were taught which involved part or all of a crew. Related tasks were tested in an exercise given when the players had completed training on the tasks. The results of these exercises show that there were 478 Go's on 629 tries for a Go rate of 76%. The areas with the smallest Go rates were target acquisition (59%), range determination by gunner (35%), thermal mode operations (61%) and the COAX and 25mm live fire exercises (51% combined).

b. Other hands-on tests (HOT)

A target acquisition and engagement (TAE) hands-on test was developed by personnel from DTD and TRASANA to assess the players' proficiency at acquiring and engaging targets. This test was given on two occasions; at the end of the training program (TAE I) and the end of the FDTE (TAE II) . The test was designed to simulate a CFV in an overwatch position observing a sector for enemy targets. Prior to beginning the exercise, the crew was allowed to observe the sector for one minute to familiarize themselves with the terrain. During the test, in which the CFV was always stationary, the assigned TC for the crew acted as TC while each crew member was tested as gunner. Then the assigned gunner acted as TC while the TC was tested as gunner. As each pair was being tested, the remaining crew members were prevented from observing the sector.

c. TAE I.

1. Each TC/gunner pair was tested for one iteration in which four targets were presented. These were full-scale targets mounted on SAAB devices which were raised and lowered by the test administrators. Two test administrators gathered the data for all the tested pairs. Both were positioned on the CFV where they could observe the soldiers being tested but located so as not to interfere with their performance during the test.

Specific items of interest are:

- The standard for target acquisition (5 sec from target exposure) was met 34% of the time.

- The target was correctly identified by type 42% of the time.

- The mean time for the gunner to determine the range to the target was 25.8 seconds.

- The mean time required by the gunner to determine the range to the target after he acquired it (8.5 sec) exceeded the time standard for acquisition plus ranging (8 sec).

- The gunners were able to correctly determine the range to the target (within 200 meters) 60% of the time.

- The target was engaged within 20 seconds (25 seconds for TOW engagements) 29% of the time.

- The proper ammo for the engagement was selected 38% of the time.

Analysis of the data by individual engagement revealed the following:

- The target was correctly identified and the proper ammo selected 27 times (29%).

- Tank targets were correctly identified 23 times (35%); TOW was selected to engage these targets 11 times (48%).

- Of the 92 total engagements, time standards for all required tasks were met seven times (8%), though all tasks were not necessarily performed correctly.

- Of the 92 total engagements, there were only two in which each required task was correctly performed and performed within the time standard.

2. The results of this test indicate that the players were not able to meet the time standards set for the tasks involved. Also, the players incorrectly identified the targets 58% of the time. Several of the players commented that the lack of a gun tube on the target silhouettes made them difficult to identify. While this may be true, the silhouettes used in the test (and the live firings) were standard threat targets for Army training. Another observation concerns the players' selection of ammo. In those cases where the target was correctly identified by type, the proper ammo was chosen 29% of the time. More specifically, TOW was selected against correctly identified tank targets less than half (48%) of the time. This represents a serious shortcoming of the training program in preparing the players to engage and defeat the enemy.

d. TAE II.

1. The players' overall performance on the second test equalled or exceeded their performance on the first test in all areas except "time to range after acquisition" and "TOW selected against correctly identified tank targets". The increase in time to range after acquisition was thought to be due in part to the lack of experience at this task on the part of the scouts and drivers. Analysis of those engagements in which an assigned TC or gunner was tested at the gunner's position shows that the mean time for these crewmen to range to the target after acquiring it was 9.3 seconds. The mean time for the remaining crewmen was 14.2 seconds. Corresponding results from the first test are 7.0 and 9.2 seconds, respectively. Both groups showed an increase in time to perform the task from the first to the second test. Thus the difference between the times was not caused solely by the crewmen being less familiar with CFV turret operations. A possible explanation for the change is that the players did not practice ranging to target within the time standard during the FDTE frequently enough to maintain the level of proficiency they had acquired during the training period.

2. The players also showed a decrease in the percentage of engagements for which TOW was selected against correctly identified tank targets from the first to the second test. There was also an overall laxity in issuing and following fire commands during the second test that was not as prevalent during the first test which could have been due to lack of structured, supervised practice in this area.

e. Results of Written Skills and Knowledge (S&K) Tests.

1. The written tests were developed by TRASANA from available training materials. These tests were not part of the plan of instruction for the taining but were given by TRASANA in an attempt to measure the players' knowledge about the CFV in specific areas.

2. The players were given three skills and knowledge (S&K) tests during the training and FDTE, (i.e., S&K I, S&K II and S&K III). S&K I was given at the end of the individual training phase; S&K II at the end of collective training and S&K III at the end of the FDTE. The tests are similar but not identical since some questions and/or choice of responses were changed, added or deleted as the players' familiarity with the CFV increased.

3. It must be noted that the high first-time "GO" rates of the class and comp HOT (98% and 96% respectively) discussed earlier reflect a greater level of knowledge of CFV tasks than do the average scores on S&K I and II. The most likely cause of this discrepancy is that the two types of tests measure different aspects of the tasks involved. The hands on tests (HOT) require the manual execution of skills and knowledge in the tasks involved while the S&K test mental recall of information about the tasks. Direct correlations between performance on the HOT and S&K tests are not meaningful since the number of "NO GO", and therefore the statistical variance, on the HOT is too small.

4. Comparing the results of S&K II and S&K III reveals that the players' scores remained virtually constant in the areas of turret operations and weapons (categories 1 and 2) while decreasing considerably in weapons maintenance. Thus, the players' knowledge of the items tested was not improved, and actually decayed in category three during the FDTE. Significant positive correlations were found between the scores on the two S&K tests in categories 1 and 2 and the total test, which indicates that the decrease in scores was not based solely on the part of those players who were less familiar with the turret and weapons system (i.e., drivers and junior scouts).

5. The results of S&K I and S&K II were also compared to the results of the target acquisition hands-on test given at the end of training (TAE I). Significant correlations were found to exist between the players' scores on categories two and three (weapons and weapons maintenance) of S&K II and their performance at ranging to the targets during the hands-on test (higher tests scores were associated with less range time). No relationships were found between the results of S&K I and the hands-on test.

f. 7.62mm COAX firings (crew subcaliber exercise).

1. Each of the fifteen players was scheduled to serve as gunner during three iterations of a day course (three targets) and a night course (five targets). Due to weapon/target malfunctions and other complications, some players traversed the course less than three times while others traversed them more often. The COAX was used as a subcaliber training device during these firings and a crew of three was on board the CFV during the engagements. The targets used were half-scale threat vehicles (except for one full scale troop target in the night course) and were set at actual distances within 1000 meters of the CFV. The CFV was moving during all engagements of the day course and stationary during the night course. Results are shown in the following, table 1.

TABLE 1

COAX ENGAGEMENTS WITHIN
TIME TO KILL STANDARDS

Conditions	Exposures	No. Times Standards Met	Percent of Exposures
Day, single target	59	27	46
Day, multiple targets	29	8	28
Night, single target	0	---------	--------
Night, multiple targets	76	7	9

2. The table shows that the time to kill standards could not have been met in more than 46% of the single target, day exposures; 28% of the multiple target, day exposures; and 9% of the multiple target, night exposures. Thus, the players did not consistently meet the standards for the firings.

g. 25mm firings (crew combat exercise).

Data from this exercise concerning accuracy of the 25mm weapon system are classified. The numbers of targets engaged and engagement times are not classified and are discussed in the following paragraphs.

1. The same fifteen players who served as gunners in the COAX subcaliber firings also participated in the 25mm live fire exercise. Targets were full scale threat vehicle silhouettes at ranges of 1000 to 1500 meters and were mounted on SAAB devices so that they could be raised and lowered by remote control. The firings consisted of three iterations for each of the fifteen players of a day and a night course. Both courses were similar to those used during the COAX exercise except that the targets were full size at longer ranges.

2. The percentage of targets engaged during the 25mm firings shows a substantial increase over that from the COAX exercise. Several factors possibly contributed to this increase in target servicing proficiency. Among these are:

- The practice received during the COAX firings provided experience which made the players more proficient at the tasks required to acquire and engage the targets.

- The targets were full size instead of the closer but half-scale targets used during the COAX exercise.

- The players were more interested in, and excited about, firing the 25mm gun than the COAX.

h. Discussion of COFT training.

1. During the individual training phase, all of the players received training in gunnery procedures using the prototype COFT*. Due to the late inclusion of the COFT into the program, this training was given primarily as a "filler" for time in which a player was not otherwise occupied. Therefore, not all players received the same amount of exposure to the COFT.

2. COFT training consisted of a series of target presentations in which the player was required to determine range and engage the target. Range determination procedures on the COFT differ from those on the CFV in the following areas:

- During COFT engagements, the player had only to announce the correct range and the instructor would press a switch which caused the COFT to adjust the impact of the rounds properly. On the CFV, the gunner must move his right hand to the range control knob and dial in the proper range. This motion could affect his ability to keep the target in the sight, especially at long ranges.

- Only three different ranges are available for COFT engagements (1200, 1500, 3000m). Unlike actual firing conditions, range to the target could easily be identified by the size of the target in the sight. The players quickly learned the relationship of target size to range and therefore were less careful to perform proper ranging techniques before announcing the range. This became evident during range determination exercises early in the collective phase, when the players tended to identify target ranges at 1200, 1500 or 3000 meters.

3. Comparison of COFT results to live fire results.

- The COFT training was conducted differently from the live firings in that the players were not restricted to time limits on the COFT. They were allowed to engage the target until it was hit (or moved off the screen for a moving target). Therefore, comparisons of mean times to engage and hit are not meaningful. The percentage of targets hit on the COFT in both the stationary firer-stationary target and moving firer-stationary target modes is much greater than that achieved during the COAX and 25mm exercises. Tests of relationships between COFT and 25mm data in both percent of hits and mean difference between hit time and engagement time for the fifteen players who participated in the 25mm exercise found no significant correlations.

- The players engaged and hit a much greater percentage of targets during the COFT training than during the COAX and 25mm live firings. This is partially due to the absence of need to acquire targets on the COFT. (Targets are readily visible on the COFT display). Another factor influencing these differences is that the players were not limited during the COFT training by the time constraints enforced during the live fire exercises.

6. ASVAB AND ATTITUDE ANALYSIS RESULTS.

a. Four ASVAB areas (OF, MM, GM and EL) and the SelectABLE scores correlated positively with scores on all three written tests. The MM scores were positively correlated with engagement times on the target acquisition hands-on test (TAE I). The higher MM scores also correlate with less time to engage after acquisition.

b. The players' responses to the attitude surveys were predominately positive. Attitudes toward the training materials, CFV night/buttoned-up operations and comfort and safety of the vehicle were less positive than those in the other areas surveyed.

c. In general, more positive attitudes toward the training were associated with better performance, as measured by the written and hands-on tests.

7. SUMMARY FINDINGS

a. The players participating in the training program were representative of the overall Army population of 19D MOS soldiers in terms of aptitude and demographics.

**This is the same COFT used during training for the IFV OT II, but is not related to the Fighting Vehicle System (FVS) UCOFT being designed for use in future IFV/CFV training programs.*

b. The players' performance on hands-on tests given by the instructors during the individual training phase indicates that the training program provided the skills necessary for successful accomplishment of the involved tasks within established time standards. Problem areas involving these tasks which arose during the collective training phase were:

°Loading the COAX and 25mm weapons

°Performing immediate action and misfire procedures on the COAX and 25mm weapons.

c. In loading the 25mm gun, the players experienced difficulty in properly correcting the ammo feed chutes. Improper connection of these chutes caused many stoppages which had to be corrected by the instructors or, in a few instances, contractor representations. The players were also unable to consistently perform the misfire and immediate action procedures in the proper sequence. The training program was not effective in promoting the retention of the skills they learned in the performance of loading and misfire/immediate action tasks. More training and/or practice time is needed in these areas to better enable the soldier to retain skills gained through the training program.

d. Written tests given during the training and at the end of the FDTE indicated that the players' knowledge of the CFV and related tasks was not as high as the results of the instructors hands-on tests given during the training would suggest. Also, scores on the two skills and knowledge tests given during the training decreased in the areas not being taught during the interim and increased in areas being taught. The practice which the players received during the conduct of the FDTE did not increase the players' knowledge of the CFV and its systems since scores on the written test given at the end of the FDTE were the same or lower than those on the same test given at the end of training.

e. The players engaged and hit a much greater percentage of targets during the COFT training than during the COAX and 25mm live firings. This is partially due to the absence of need to acquire targets on the COFT. (Targets are readily visible on the COFT display). Another factor influencing these differences is that the players, during the COFT training, were not limited by the time constraints enforced during the live fire exercise.

f. Comparisons of results from the COFT training and the target acquisition and engagment hands-on test given at the end of training showed no significant difference in mean time to determine range to the target (after gunner acquisition for the hands-on test) between the two. A significant difference was found in mean time to engage the target (after gunner acquisition for the hands-on test) with the players engaging targets more quickly during the hands-on test. This could be the result of inexperience on the part of the players when the COFT engagments used in the comparisons were conducted. A signficant positive correlation was found between mean engagement times of COFT targets and corresponding times from the hands-on test.

g. During the COFT training, all COFT vehicular targets were engaged within the 25mm weapon system, even at ranges of 3000 meters. The failure of the instructors to correct this situation probably contributed to the players' training to select the 25mm for engaging targets, including targets identified as tanks, during the target acquisition and engagement hands-on tests.

Privacy Act and the Data Base: Implementation of the Privacy Act

William B. Camm, Staff Assistant for Tests, Technical Information Division, US Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia 22333

The legal constraints of the Privacy Act and the increased legislative pressure to handle greater amounts of personal data faster and better pose many new problems for social science research in the US Government. The immediate, intermediate, and long-range solutions to the dilemma can be achieved through the development of a precautionary, systematic set of collection and storage procedures; full use of comprehensive data bases; and the insulation of each contributing set of data.

This paper is offered as a contribution to the current discussion on the legal constraints of the Privacy Act of 1974 (P. L. 93-579) on the pressing requirement that the Department of Defense handle greater amounts of complex personnel research data faster, better, and at minimal cost. The mandate to minimize the cost of collecting, maintaining, and using personal data and, at the same time, maximize the utility of the collected data is articulated in the Paperwork Reduction Act of 1980 (P. L. 96-511). The act became effective 1 April 1981.

The technology capable of cogent management of information resources is here, and we already know quite a bit about how to apply it in terms of cutting costs, enhancing usefulness, coordinating and sharing common procedures, and improving service to management and the user.

The problem, however, is to insure that the collection, maintenance, use, and dissemination of personal data and information are consistent with the Privacy Act. The restrictions and limitations imposed by the Privacy Act loom large as a potential hindrance to effective information resource management in terms of information control, resource contraints, cost of data, added hardware, and special computer programs. To date, neither the impact of the new wave of information resource management on the Privacy Act nor the constraints of the Privacy Act on the Paperwork Reduction Act has been sorted out. The Office of Management and Budget (OMB) has been assigned the responsibility for providing overall direction in the development and implementation of policies, principles, standards, and guidelines in all areas of P. L. 96-511. Privacy Act enhancement is on the OMB agenda for April 1983.

All Federal agencies are moving ahead with their own interpretation of P. L. 96-511 with the expectation that the resulting implementation will be found acceptable. The Office of the Assistant Secretary of Defense, Information Control Division, has established DOD-wide policy to insure compliance with P. L. 96-511. The Army has established an Information Management Office under the Office of the Chief of Staff to define, develop, and manage the Army information resources program.

Organizations that deal with personal information have reached informal agreements on most aspects of the program but have avoided special problems with regard to processing and maintaining personal data--especially with the concept of an integrated data base. Nevertheless, the problems of effective information resource management and protection of individual privacy are quite real and very pressing and cannot be ignored.

THE PRIVACY ACT

The Privacy Act is imposed on executive departments, military departments, Government and Government-controlled corporations, other establishments in the executive branch including the Office of the President, and independent regulatory agencies. Congress and its agencies (e.g., GAO) are exempt. So are Federal Courts. It limits the manner in which they collect, use and disclose information about people. The act was codified as 5 USC 552a in 1976. The act gives the individual the right to be protected against the power of officials with access to data banks. There are three related aspects to the Privacy Act rights:

> Personal autonomy--the right to make a choice about personal behavior and lifestyle.
>
> Freedom from outside interference--the right to be left alone.
>
> Protection of private information--the right to control where and how information about oneself is communicated to others.

This portion of the paper focuses on the third point, control and protection of personal information.

Since 1965, personal privacy has become an important social value, covered under tort laws, in the United States. Privacy is related to personal freedom and, although rights of privacy are not expressly mentioned in the Constitution, is supported by the Supreme Court's language used in most of its important decisions. The constitutional amendments most commonly cited in this regard by the Supreme Court are the first, third, fourth, fifth, ninth, and fourteenth.

In 1976, an inventory of Federal data systems revealed that 97 agencies had a total of 7,000 records systems containing nearly 4 billion dossiers. The Department of Defense alone had 2,219 systems with 321 million different names and records (OMB, 1976). Most of the records systems at that time were not a matter of public record. The Privacy Act prohibits secret files and further states that individuals should be able to find out what information about them is contained in Federal records and how that information is used. For example, a person is able to prevent personal information that was given for one specific purpose from being used for another purpose without his or her consent. Provisions will be made for the individual to correct and amend personal records in possession of the government.

Government agencies handling identifiable personal data should show that such data are reliable and current and take positive steps to prevent their misuse. Collected data should also be safeguarded and securely stored if they contain identifiable information. For research use, the connection between the names and data should be destroyed when no longer needed. Code numbers and code words can be used if several sets of data are collected on the same person. A number of methods for storing personal data are described in the literature (Boruch, 1971a, b). If knowledge of illegal activities is requested, anonymity should be guaranteed so that the data cannot be subpoenaed in legal proceedings. Insulated data banks might be considered. Research data are not automatically privileged information. There are a few exceptions, such as data regarding drug research. Congress and courts may, and often do, subpoena such data.

The success of the Privacy Act is hard to measure objectively. The enforcement of data protection regulations and the supervision and control of the collection and storage of information about individuals depend, for the most part, on the good faith of the agencies and legal action by individuals. Congress believed that self-regulation was the best initial method for control because it eliminated the need for an additional government agency and, at the same time, would aid the necessary advance in the technology of information collection and storage. Control agencies were not to be considered unless the agencies themselves proved that self-regulation had failed. The potential for frustration of the law is so great that the Privacy Protection Study Commission (1971) recommended that the Privacy Act be broadened to include all items that an agency can readily identify in all of its systems to insure compliance with the Privacy Act. So far, Congress has not implemented the Commission's recommendation.

Violations of the Privacy Act are misdemeanors subject to a maximum fine of $5,000. Unlike damage actions brought against an agency, criminal penalities are imposed on the person who committed the crime. The punishment, if there is a conviction, is applied to any

> Agency officer or employee who knowingly or willfully makes improper disclosures of information pertaining to an individual.
>
> Agency officer or employee who willfully maintains records without meeting Notice Requirements Requests (i.e., maintains a secret system of records).
>
> Person who knowingly and willfully requests or obtains individual records from an agency under false pretenses.

If the court finds that an agency (its officers or employees) acted in an intentional or willful manner, the complainant may receive actual damages ($1,000 minimum). But it is difficult, in most cases, for the complainant to show proof of intentional and willful agency misconduct. The complainant must also show that the conduct was greater than "gross negligence"; "ordinary negligence" on the part of the agency does not meet requirements of the law as it is written. In addition, the complainant must also prove actual damages by establishing that the agency's action had a direct adverse impact upon him or her. Finally, if the individual wins, the U.S. Treasury (not the agency or its members) is liable for the actual damages, court costs, and attorney fees. This situation tends to dampen the deterrent effect that civil actions may have upon data collection practices of agencies (Bushkin & Schaen, 1975).

THE DATA BASE

A data base may be viewed as a digital computer version of a manual file system. The manual file system comprises file folders identified by a name or number. The computer file consists of records, each identified by a primary key and secondary keys, for example, name, age, rank, and Social Security number. At this point, the computerized record system departs from the manual system. Access to the items in the computerized system can be made through the primary or any secondary key, or through any other indicator in the individual record. Users of computerized records systems are often in remote locations, and restrictions, like code names for the primary key or identification tab of a single system, no longer exist.

The recent trend, under the impetus of the Paperwork Reduction Act, is toward integrated data bases where a collection of data or records is linked together using a common identification key. The reason for the innovation is related to a greater need for individualized information and a growing proficiency in processing and interpreting data. Also, as expected, data collected for one purpose is frequently useful for related purposes.

At this point, the distinction between records that relate to individuals for the purpose of taking some sort of action concerning that individual and records that are collected and maintained for the purpose of planning and policy decisions should be made. The former, in the strict sense, is termed a system of records; the latter is statistical record. However, most records are mixed, and it is rare to find a true statistical record in either Government or academic research. A true statistical data base cannot contain information that can be related to an identified individual, and no individual contributing to the data base should be identified with it.[1]. The Army Research Institute for the Behavioral and Social Sciences (ARI) collects and maintains systems of records until such time as the data are edited, coded, stripped of the personal identification, and entered into the data base. The ARI Systems Notice (ARI, 1980) covers, at this writing, all ARI systems of records of the moment and the future, provided the data collection effort remains within the operational confines of the public notice. If a new and different system of records is contemplated, then an additional notice, or modification of the current notice, will be required. The new notice must be published in the Federal Register at least 90 days prior to any data collection for the new system of records.

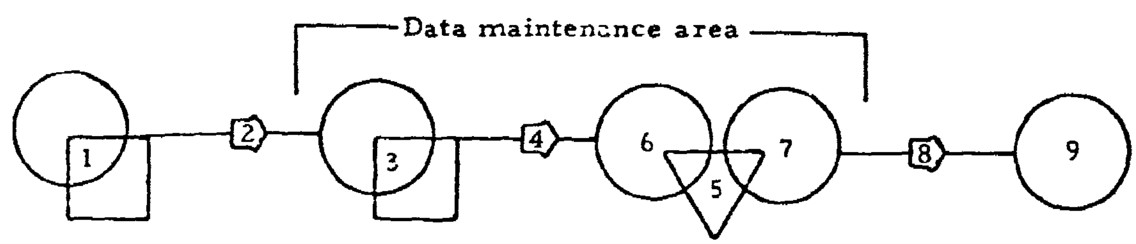
DATA COLLECTION AND STORAGE PROCEDURES

The procedure involved from the start of the data gathering through the final destruction of the system of records (i.e., removal of personal identifiers) and the publication of the results for the various users may theoretically be compromised at a number of points during the collection, transmission, storage, and processing of the data. Nine arbitrary points are conceptualized here for the purpose of illustration in Figure 1.

The data collection point 1 surveys, questionnaires, tests, interviews, or ratings is obvious and frequently overlooked despite the Privacy Act Statement at that point stating that "Full confidentiality of the responses will be maintained in the processing of the data. . . ." (DA Form 4368-R). The Privacy Act requires that all agencies involved in data collection--in the development of a data base there may be several--provide

[1]Insofar as the Privacy Act is concerned, however, the only operative criterion is whether or not the agency does in practice retrieve the information by reference to some personal identifier.

appropriate administrative, technical, and physical safeguards. The common threat to personal information at point 1 is the person who is authorized to have access to the information for one purpose but who misuses that same information for an unauthorized purpose. The entire data collection operation, if possible, should remain under a single work group. It is tempting for the researcher to ignore most of the problems at point 1 and go on to the second potential compromise point (transmission). The personal information is easiest to protect in the computer-based area. The transmission of the data (point 2) may be by messenger, mail, telephone, or microwave and is subject to compromise during transmission and upon receipt. Any privacy compromise here is seldom intended and is most likely the result of careless handling. Security compromise during transmission is not specifically treated in this paper. The editing and coding process (point 3) is the first step in preparing the data for the computer and is the time to check for accuracy, relevance, timeliness, and completeness. It is also a good time to remove the personal identification in preparation for linkage with additional information in the integrated data base unless that linkage is necessary for the subsequent interpretation of the data. Data transmission (point 4) to the computer area is usually less of a risk than point 2. However, the information must be checked, edited, and sorted and may easily be identified by resourceful people. Points 5, 6, and 7 involve checking the processing of the data being edited and stored. During format checks of tables, graphs, and the like, careless handling may result in compromise. Error listings are another source of compromise at this point. The location of each item of information should be recorded and confined to the computer area; extraneous data should be destroyed when no longer useful. Finally, point 8 is transmission of the data (the report) to the user, point 9. Exploitation can occur when common and unique properties of individuals are displayed in the reports. It is then a simple matter to sort, count, and identify individuals and/or groups from the final report. For example, tabulation of results may yield grade level, age, sex, location, and other properties that with cross-tabulations identify individuals and/or groups.



1. Data Collection
2. Data Transmission
3. Editing and Coding
4. Data Transmission
5. Data Preparation
6. Computer Processing and Storage
7. Tabulation and Display of Results
8. Data Transmission
9. Report: The results to user

Figure 1. Flow from personnel information initial collection to statistical record to final report. Numbers represent potential compromise points.

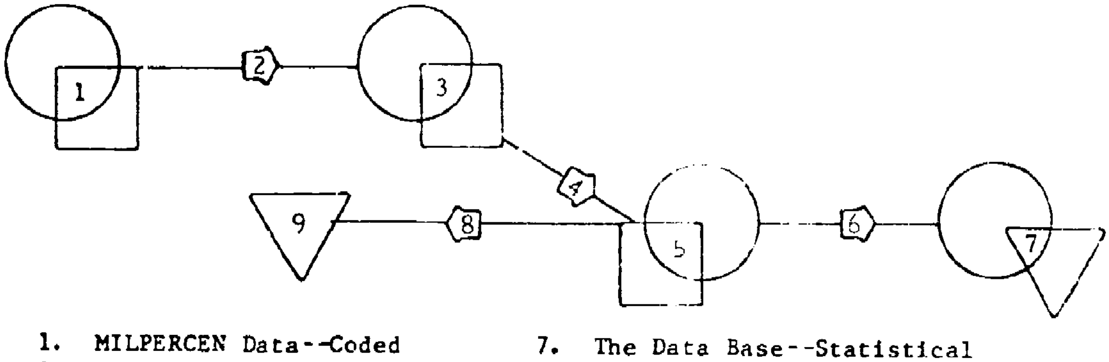
In practice, the situation is more complicated. Longitudinal studies which involve collecting and maintaining information over a period of time may present problems. A statistical data base of this sort needs an insulated method of linking recent data with data already stored. To complicate matters, a secondary user or users are often involved. And most problems arise--at least insofar as privacy safeguards are concerned--when the primary user establishes the data base for administrative purposes and the secondary user is more interested in research, or vice versa. Often, there is no relationship of purpose between the records system of one user and the established data base of another.

PRIVACY SAFEGUARDS

Privacy safeguards for data bases are similar to those required for most records systems. Certain data bases, for example those concerned with current sensitive issues, such as medical histories, performance by ethnic groups, illegal actions or country of origin (Barnes, 1979), are subject to intentional invasion for several reasons by individuals whose interests range from apprehension concerning possible misuse, real or imagined, of the information contained in the data base to intelligence-gathering activities of foreign governments. Added precautions might be considered.

For example, the data from MILPERCEN's proposed data base are coded with a cryptographic code known only to MILPERCEN. The coded data plus identifying information are sent to ARI to merge with ASVAB data, which is also coded using the identifying information to link with the MILPERCEN record (Figure 2). The identification is then deleted. The merged file is given to ARI's Personnel Utilization Technical Area. MILPERCEN cannot obtain anything other than their own data from the file, and ARI cannot meaningfully identify data from MILPERCEN but will have the necessary information for a validation of ASVAB. The same scheme can be used in longitudinal studies with different independent groups. The code linkage can either be destroyed or stored in a safe place beyond the reach of all but extraordinary requests.

Assuming reasonable precaution in data collection, maintenance, storage, and reporting, the insulated data base with its disposable code links and the resulting statistical record will easily meet future requirements for privacy protection of ARI integrated data bases. There are many other effective methods to insulate and link record systems. There is no one best way to protect personal information. The point is that such protection can and should be provided.



1. MILPERCEN Data--Coded
2. Data Transmission
3. ARI ASVAB Data--Coded
4. Data Transmission
5. Merge Data--Edit and Match Codes
6. Data Transmission of All Coded and Merged Data
7. The Data Base--Statistical Records only
8. Transmission of Cryptographic Key
9. Safe Storage of Code Key

Figure 2. Schematic flow and proposed development of one insulated data base.

Army Research Institute. ARI Systems Notice. A1306.01 DAPE. Federal Register, 45FR No. 223, 75736. Nov. 17, 1980.

Barnes, J. A. Who Should Know What? Cambridge, England: Cambridge University Press, 1979.

Boruch, R. F. Assuring confidentiality of responses in social research: A note on strategies. American Sociologist, 1971a, 6, 308-311.

Boruch, R. F. Maintaining confidentiality of data in education research: A systematic analysis. American Psychologist, 1971b, 26, 413-430.

Bushkin, A. A., & Schaen, S. I. The Privacy Act of 1974. McLean, Va.,: System Development Corporation, 1975.

Office of Management and Budget. The first annual report to the President for CY 1975. Washington, D.C.: U.S. Government Printing Office, 1976.

Privacy Protection Commission. Personal privacy in an information society. Washington, D.C.: U.S. Goverment Printing Office, 1971.

TRAINING AND COST EFFECTIVENESS OF TWO AVIONICS MAINTENANCE TRAINING SIMULATORS[1]

Louis F. Cicchinelli, Ph.D.
and
Kenneth R. Harmon
Denver Research Institute
Denver, Colorado

Summary. Two studies were conducted to determine the relative training and cost effectiveness of two simulated versions of the F-111 Converter/Flight Control System maintenance test station as compared to the operational test station equipment. Study 1 was designed to assess classroom and field performance of students trained on the actual test station and those trained on the 6883 3-dimensional simulator. Further, the life cycle costs of using the two systems for training were compared. The results of Study 1 indicated that there were no significant differences in either classroom or field performance as a function of training equipment. However, the comparison of system costs indicated that actual equipment costs were approximately twice those of the 6883 3-dimensional simulator when calculated over a 15-year life cycle. Study 2 was designed to make similar performance and cost comparisons among the 3-dimensional simulator, a 2-dimensional simulator, and the operational test station equipment. The preliminary results of this study show that there were no significant differences in student performance as a function of the training device employed. The life cycle cost commparison among all three training devices is not yet complete.

Introduction

Traditionally, hands-on training in the maintenance of the F-111 Converter/Flight Control System has been provided to avionics maintenance trainees using operational 6883 test station equipment. In fact, the use of operational test station equipment has been standard procedure throughout the Air Force aircraft maintenance training schools. However, in the early 1970s, the need for more cost effective trainers became apparent. Miller (1974) and Miller and Gardner (1975) provided an extensive analysis of the need for simulated training in general and, specifically, for the development of the 6883 3-dimensional (3-D) simulator employed in part of this evaluation. Factors influencing the decision to proceed with the development and implementation of 6883 simulators included the high cost and low reliability of actual equipment trainers (AET), safety factors, high noise levels in AET work areas, and the limited scope of training that can be accomplished on AETs because appropriate malfunctions cannot be inserted without extensive and costly modifications.

Unfortunately, methodologically sound comparative studies of the training and cost effectiveness of simulators and actual equipment used for training are conspicuously rare and available literature on the use of simulators specifically for maintenance training is even more limited. Of the studies

---

[1]This work is being conducted by the Denver Research Institute for the Air Force Human Resources Laboratory at Lowry AFB under contract No. F33615-78-C-0018. A more detailed discussion of Study 1 is provided in report number AFHRL-TR-80-24.

which have been conducted, two important points have been established: maintenance training simulators can be cost effective (Miller, 1978; Fink, Shriver, Downing, & Miller, 1978; Montemerlo, 1977), and maintenance training simulators can provide training at least comparable to the actual equipment (Miller, 1974; Hurlock & Slough, 1976; Daniels, Datta, Gardner, & Modrick, 1975).

While only a few simulator maintenance training programs have been in operation long enough to assess implementation effects, some problems have become apparent. First, there is the need to assess the fidelity of the training device in order to establish the equivalency of training contexts. Otherwise, the evaluation design may incorporate different training and testing contexts and render the results inconclusive. Special attention must be given to psychological fidelity since the level of realism can have important impacts on training effectiveness. Second, there is the need for a criterion measure of training effectiveness. Previous researchers have had difficulty in establishing such a measure without incurring high costs (Miller, 1978). And finally, problems can arise as a result of instructor opposition. Instructors may see their teaching role threatened by the incorporation of simulation methods or may disagree with required changes in the academic structure.

The present studies necessarily involved some of these issues. It was important to consider the role of the simulated devices in training, their fidelity to AETs, and user acceptance of the new equipment. Furthermore, the cost analysis of the simulators and AET station forced examination of concepts and assumptions that link training effectiveness with the respective system costs.

Study 1

The overall objective of Study 1 was to design and implement a comprehensive comparative cost and training effectiveness evaluation between the 6883 3-D simulator and the operational 6883 test station equipment.

Research Design

This study was designed to examine three major questions: (1) the relationship between mode of training and classroom performance, (2) the relationship between mode of training and job proficiency in the field, and (3) the relative costs of using the two devices for maintenance training. The basic design used to assess classroom and field performance is shown in Figure 1.

For the purpose of assessing classroom performance, four experimental groups were defined by the two training modes and two testing modes (Groups A-D). Clearly, performance differences in operating and maintaining actual test stations as a function of the training equipment was of primary interest. However, two testing modes, actual and simulator equipment, were used because it was expected that the simulator might provide training or testing capabilities which were not available on the actual equipment. Further, it was necessary to determine the extent to which any observed differences in performance were due to familiarity with the test equipment.

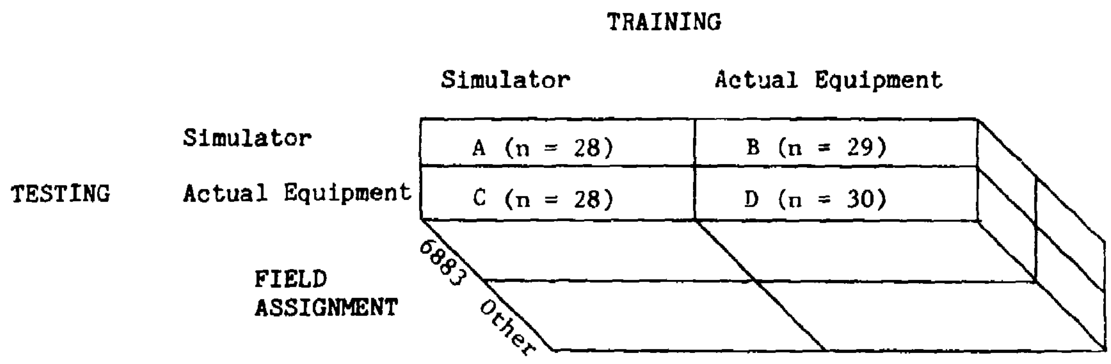


Figure 1. Research Design Used for Comparing Field Performance.

As shown in Figure 1, to assess the impact of training mode on job performance in the field, it was necessary to consider eight experimental groups. That is, the assessment of field performance was conducted in view of the four levels of training resulting from the various combinations of classroom training and testing modes.

The comparison of costs associated with using the 6883 actual test station and the 6883 3-Dimensional simulator for training was based on the "ingredients approach" discussed by Levin (1975) in which cost elements are identified and evaluated consistent with the ATC acquisition and training environment. The cost elements or ingredients associated with each cost category were evaluated either as one-time costs (primarily Investment Costs) or as Recurring Annual Costs, consistent with the Air Force perspective on economic analysis (Williams, 1977). The major cost categories used in the present analysis were: facilities, equipment, instructional material/training, personnel, students, and miscellaneous. For purposes of the cost analysis, it was assumed that both trainers had equal training effectiveness and that the life cycle cost comparison between trainers would indicate which device was the most cost effective, i.e., the trainer exhibiting the least total cost of ownership. This was a useful approach since it established baseline data and was consistent with the original simulator design objective of developing a functional (training) replacement for the 6883 test station (Miller & Gardner, 1975).

Methodology

A total sample of 115 F-111 avionics maintenance trainees participated in Study 1. Students were assigned to the treatment groups defined by training sequence and the two levels of training mode and test mode. Although training sequence was initially considered in the assignment strategy, subsequent analysis indicated that performance did not vary as a function of training sequence. The distribution of students among the four resulting experimental groups is shown in Figure 1.

The assignment of students to groups was essentially random. Tests for aptitude and prior achievement indicated no biases across groups. Further, there were no apparent biases in assignment with regard to student gender.

One day at the end of the Converter/Flight Control Systems practical block of instruction was made available for data collection. This temporary departure from the normal training schedule, effective for the duration of this project, allowed data to be collected without altering the usual 6883 training protocol. In order to collect all information required from an entire class of students (usually six airmen) in a single day, it was necessary to administer the performance tests and conduct the interviews in a different order for each student. Administration of the trouble-shooting performance test for the first student was followed by the personal interview. During the time that any one student was being administered the trouble-shooting performance test, the remaining students, who had either completed or were awaiting their turn to be tested, were given a Projected Job Proficiency Test. When all students in a class had completed the performance test, the Projected Job Proficiency Test, and had been interviewed, the evaluation day was complete. In addition to this primary data collection effort, other performance data was obtained from the student records. Finally, subsequent to placement in the field, students were again interviewed concerning their training experience and supervisors were asked to rate student field performance.

Data needed for the cost analysis were obtained primarily from Air Force acquisition and maintenance records. However, it was also necessary to monitor training deviations and unscheduled maintenance of equipment throughout the study to insure an accurate estimate of operating costs of both systems.

Results

A central issue to be addressed by this investigation was whether simulator and actual equipment training would result in equal levels of student performance on a practical trouble-shooting problem as might be encountered in the field. The test which was developed consisted of a timed, serial, hands-on, 29 item task. The test allowed three overall dependent measures of performance: a total score, the total time necessary for test completion, and a three point rating of the degree of assistance required for test completion. Table 1 shows the mean scores for each of the experimental groups on each of these three measures.

TABLE 1

Overall Trouble-Shooting Test Proficiency as a Function of Training and Testing Modes

			Training Mode	
	Test Mode		Actual TS	Simulator
Total Test Scores	Actual	N	24	28
		$\bar{X}$	23.56	22.82
		σ	1.86	1.47
	Simulator	N	29	28
		$\bar{X}$	22.66	22.86
		σ	1.56	1.53
Total Time in Minutes	Actual	N	24	18
		$\bar{X}$	50.21	53.75
		σ	[illegible]	7.72
	Simulator	N	29	28
		$\bar{X}$	58.10	55.93
		σ	7.81	9.66
Instructor Assistance Required	Actual	N	24	28
		$\bar{X}$	2.24	2.29
		σ	.66	.85
	Simulator	N	29	28
		$\bar{X}$	[illegible]	2.29
		σ	.64	.71

Completion times for this actual/actual (training mode/test mode) versus simulator/actual comparison did not differ significantly ($F(1,105) = 2.193$, $p = .14$) while there was a marginal difference in total test scores ($F(1,105) = 2.903$, $p = .09$). That is, actual-trained students scored slightly higher on actual equipment testing than simulator-trained students. This effect is minor, though, and may be the result of machine-specific experience rather than qualitative differences in training. It seems entirely plausible, for example, that students would be more nervous being tested on the real test station if their only experience had been with a simulator where the ramification of errors are not as serious.

A brief examination of the degree of assistance required was carried out despite this measure's lack of sensitivity. The group means reported in Table 1 are extremely similar, and a chi square analysis showed no evidence that ratings were related to experimental conditions ($X^2(6) = 9.63$, $p = .14$).

In sum, then, simulator- and AET-trained students did not differ appreciably with respect to overall trouble-shooting ability as measured by the practical test. A very slight advantage in test accuracy was found for actual- as opposed to simulator-trained students tested on the actual 6883, but this finding was not mirrored using commpletion time as a measure. The simulator proved to be a somewhat slower testing device which caused some interference for actual-trained students.

The Projected Job Proficiency Test, a multiple answer performance test, also indicated no significant differences among groups as a function of training mode. Hence, it can be concluded that all students, regardless of training mode, acquired equivalent job-related experience. In an effort to assess user acceptance of the simulator as an alternative training device, both students and instructors were asked about the utility and adequacy of the equipment. While the feedback was mixed, generally students indicated that, regardless of the equipment used, the training experience was both adequate and enjoyable. On the other hand, interviews with a limited number of instructors revealed a concern about the slow response time of the 3-dimensional simulator and identified the need for more sophisticated courseware.

The results of the life cycle cost comparison between the 3-dimensional simulator and the actual equipment are summarized in Table 2. From Table 2 it can be seen that when both Investment Costs and Recurring Annual Costs are considered, the actual equipment is more than twice as expensive to acquire and operate as a training device. Similarly, even if Investment Costs are considered as sunk costs, the cost in constant dollars of operating the actual equipment is still about twice that of the simulator $3,336,150 and $1,588,020 respectively.

In considering the policy implications of a 2 to 1 cost effectiveness ratio in favor of the 6883 simulator, it is important to realize that the estimate for the AET is extremely conservative. The costs of the CENPAC computers were not allocated to the 6883 test station and no cost element was included in the equipment category that reflects the cost of installation and start-up of the AET. Installation and start-up costs for the simulator were fully allocated and these costs were perhaps unexpectedly high due to problems encountered in bringing the simulator on-line.

TABLE 2
The Life Cycle Cost Comparison

Cost Categories	Simulator	AET
Facilities $ 110,650	$ 110,650	
Equipment 1,594,330	4,902,140	
Instructional/Materials	26,000	27,890
Personnel 94,250	72,530	
Students 357,770	357,770	
Miscellaneous	0	0
TOTAL	$2,183,000	$5,470,980
NPV (1978)	$1,501,090	$3,895,680
$/student-hour	$348 = $23/ 15 yrs student-hr	$902 = $60/ 15 yrs student-hr

Study 2

To more fully explore the use of alternative training devices, a 2-dimensional 6883 simulator was designed and developed. This training simulator is significantly different from the 3-dimensional simulator evaluated in Study 1 in a number of important ways. First, the 2-dimensional device is composed of four components, or part task trainers (PTT), which can be used for instruction in either an integrated system approach of in an individual component approach. The four components are a logic trainer, a switching complex trainer, an O-scope trainer and a flat-panel trainer. Secondly, the design of the 2-dimensional simulator incorporates less physical fidelity to the actual equipment than the 3-dimensional simulator. Finally, the switching complex trainer, a component of the 2-dimensional system, provides training which is not available on the 3-dimensional simulator and which has previously been taught only on the actual test station equipment or using more traditional classroom methods (e.g., black board).

The purpose of Study 2 was to incorporate this 2-dimensional simulator into the training and cost comparison conducted in Study 1. Initially, it was expected that the performance and cost data relative to the 2-dimensional simulator could be easily integrated with that for the 3-dimensional simulator and actual equipment already collected in Study 1. However, it became apparent that this would not be possible. Concurrent with the delivery of the 2-dimensional simulator, several course objective and training format changes occurred which resulted in dramatic modifications to the Converter/Flight Control System blocks of instruction. Thus, it was necessary to collect performance data simultaneously for all three training devices.

Since previously collected performance data could not be utilized in Study 2, a decision was made to use this opportunity to modify the data collection instruments used in Study 1 and to incorporate a preassessment test package into Study 2.

Research Design

Figure 2 shows the research design employed to compare student performance across the three training devices.

Testing Mode

	AET	3-D	2-D
AET	A		
3-D	B	E	
2-D	C		F
3-D/PPT	D	(X)	(Y)

Figure 2. Study 2 Research Design

Since the 2-dimensional simulator was developed with peripheral part-task trainers (PTT), one of which was used to provide patch-panel training for 3-dimensional simulator training, the 3-D/PPT training mode was added to the design. In sum, the Study 2 evaluation effort was designed to analyze only six of the twelve possible configurations due to a limited anticipated supply of students during the course of the evaluation. Also, since subsequent equipment malfunctions necessitated reassignment of experimental groups, not all matrix cells could be filled as anticipated, and two additional cells were included (X and Y in Figure 2), although sample sizes in these cells remain insignificant.

Methodology

A total of 119 students have been tested to date in Study 2. A major effort was devoted to redesigning student performance, user acceptance and equipment monitoring data collection forms based on observations made during Study 1. A preassessment test package consisting of three aptitude tests was also included in this study and was administered to all students at the beginning of the 6883 instructional block. This preassessment, including the Delta Concealed Figures, the Delta Reading Vocabulary Test, and the Ship Destination Test was designed to assess the spatial, verbal, and logical abilities of students. The purpose of this package of tests at the initiation of 6883 training was to ascertain if differences in student aptitude exist which could influence subsequent performance evaluation.

The two major Study 1 performance tests, the Hands-On Trouble-Shooting Test and the Projected Job Proficiency Test, were redesigned to facilitate a more rigorous item analysis and to reduce the overall amount of time needed for data collection. An additional Paper and Pencil Trouble-Shooting Test was incorporated into the performance measure package to assess students' knowledge of the electronics system and their ability to follow TO diagrams in the identification of electronic failures and malfunctions. The Student Interview, Instructor Survey and Training Monitoring forms were also redesigned to simplify implementation. The field follow-up forms remained essentially the same as those used in Study 1.

Results

Study 2 analyses are preliminary at this time since the contract period will not expire until December 30, 1981, and data collection will continue through mid-October 1981. Table 3 shows the means for the four groups on preassessment measures.

TABLE 3

Preassessment Measures for the Four Experimental Groups

Training Mode

		AET	2-D	3-D	3-D/PPT	Total
Hidden Figures	x	5.60	5.58	4.58	4.62	5.20
	(n)	(35)	(36)	(33)	(13)	(117)
Vocabulary	x	26.49	24.03	27.79	26.92	26.15
	(n)	(35)	(36)	(33)	(13)	(117)
Ship's Destination	x	35.51	36.61	35.27	33.83	35.61
	(n)	(35)	(36)	(33)	(12)	(116)

A preliminary analysis indicated that there were no significant differences in the preassessment scores among the four experimental groups. Specifically, when all four levels of training were included in a series of one-way analyses of variance, no significant differences among means were found for any of the preassessment tests. However, when the 3-D/PPT group was ignored a marginally significant effect of training mode was found for the vocabulary measure ($F(2,101) = 2.60$, $p = .079$). Table 4 shows the means for the four groups for each of the three performance measures used.

TABLE 4

Performance Measures for the Four Experimental Groups

Training Mode

		AET	2-D	3-D	3-D/PPT	Total
Hands-on Score	x	21.50	22.53	21.74	21.77	21.91
	(n)	(34)	(34)	(35)	(13)	(116)
PJPT Score	x	20.77	20.58	20.85	21.23	20.79
	(n)	(35)	(36)	(35)	(13)	(119)
P&PTS Score	x	19.29	19.47	17.52	21.23	19.11
	(n)	(31)	(36)	(29)	(13)	(109)

When all four levels of training were submitted to one-way analyses of variance, no significant differences among means were found for any of the three measures. When the 3-D/PPT group was not included in the analysis, a marginally significant effect of training was found for the hands-on test scores ($F(2,100) = 2.51$, $p = .087$). However, it should be stressed that this analysis collapsed student groups across testing modes and that being trained and tested on different systems may result in confounding of performance measures due to unfamiliarity with the equipment used for testing. In fact, an analysis involving only those groups of students who were trained and tested on the same equipment revealed no significant differences in performance on any of the three measures.

Unfortunately, at this time the data required to compare performance as a function of training equipment, when only the actual equipment is used for testing, are unavailable. This information will, of course, be a major component of the final analysis of Study 2 data.

Finally, while the life cycle cost comparison among the three training devices has not been completed, it is clear that the 2-dimensional simulator will be significantly more cost effective than the actual equipment. However, the relative cost effectiveness of the two simulated test station systems is not known at this time.

Discussion

Based on the results of Study 1 and the preliminay analysis of Study 2 data, it can be concluded that students trained on the 6883 3-D and 2-D simulators performed as well as students trained on actual equipment. It could be argued, however, that the primary benefits of employing simulated trainers were simply not realized. It is often assumed that simulators designed to replace the more costly, less reliable, and more dangerous actual equipment trainers must maintain a high level of physical and psychological fidelity. This assumption stems from the fact that the simulators are usually integrated into existing curricula and are generally used by instructors in a manner identical to the way the actual equipment had been used. Given this limited perspective, it is not surprising to observe equivalent student performance when employing simulated and actual equipment trainers. Hence, the outcome of cost comparisons is likely to become the major factor directing future procurement decisions.

It would seem that a decision to use simulators as supplements to actual equipment trainers would allow more flexibility in their design, since the actual equipment would be retained to insure compliance with existing training requirements. If simulated and actual equipment trainers are used in conjunction, improvement in performance must be demonstrated to justify the obvious additional cost of acquiring and maintaining two training devices.

In addition to performance and cost considerations, other factors may play a major role in the decision to employ simulated test stations as trainers. For example, it is unlikely that a simulator of any quality will be accepted into existing training curricula if it is not somewhat consistent with established instructional practices. To encourage instructor acceptance, the simulator should be effective as both a visual aid and demonstration tool. This would allow the simulator to be effectively incorporated into training segments (e.g., theory familiarization) which do not include extensive practical trouble-shooting experience. Such a dual purpose simulator would be almost essential if total replacement of existing equipment is planned. The potential impact of simulator training on student performance may best be acheived if a "utilization strategy" is designed to accompany the placement of the equipment into an existing training environment. Designing such a plan to highlight the real and potential uses of the simulator would insure that its unique training capabilities were tapped, and that benefits such as improved student performance, consistent training, reduced training time and cost savings might be more readily observed.

The generalizability of the findings presented here is, of course, limited. While every effort was made to adapt experimental design principles to this natural setting, it was not possible to rely on many of the premises of basic learning theory. Until parameters such as course content, training method, and duration of training, all known to affect learning, are subject to more careful control, a rigorous cost effectiveness analysis of simulation training is not possible. To answer the question, "Do simulators provide more cost effective training than actual equipment trainers?" we must be able to maximize the capabilities of simulators. Simply stated, operational test stations were not designed for training purposes, but simulators can and should be designed solely for that purpose.

REFERENCES

Daniels, R.W., Datta, J.R., Gardner, J.A., & Modrick, J.A. Feasibility of automatic electronic maintenance training (AEMT). Volume I--Design development and evaluation of AEMT/ALQ-100 demonstration facility. Warminster, PA: Naval Air Development Center, May 1975.

Fink, C.D., Shriver, E.L., Downing, D.I., & Miller, G.G. Simulators for maintenance training: some issues, problems, and areas for future research. AFHRL-TR-78-77, AD-A064692. Lowry AFB, CO: Technical Training Division, Air Force Human Resources Laboratory, January 1978.

Hurlock, P.E., & Slough, D.A. Experimental evaluation of Plato IV technology; final report. San Diego, CA: Navy Personnel Research and Development Center, August 1976.

Levin, H. Cost-effectiveness analysis in evaluation research. In Handbook for evaluation research by M. Guttentag & E. Struening (Eds.). Beverly Hills, CA: Sage Publications, 1975.

Miller, G.G. Some considerations in the design and utilization of simulators for technical training. AFHRL-TR-74-65. Lowry AFB, CO: Technical Training Division, Air Force Human Resources Laboratory, 1974.

Miller, G.G. New directions in maintenance simulation. Lowry AFB, CO: Technical Training Division, Air Force Human Resources Laboratory, 1978. (ERIC Document Reproduction Service No. ED 154 168.)

Miller, G.G., & Gardner, E.M. Advanced simulator performance specification for an F-111 test station. AFHRL-TR-75-70. Lowry AFB, CO: Technical Training Division, Air Force Human Resources Laboratory, 1975.

Montemerlo, M.D. Maintenance training devices: data and directions. In Proceedings of the 1977 Summer Computer Simulation Conference, Chicago, IL, July 18-20, 1977.

Williams, R. Economic analysis: a useful R&D management tool. Department of the Air Force, November 1977.

PAGES 279 - 288 LEFT INTENTIONALLY BLANK

HOW TO IMPROVE THE VALIDITY OF PERSONALITY QUESTIONNAIRES

WILLEM CLAEYS, PROFESSOR, HEAD OF THE DEPARTEMENT OF DEVELOPMENTAL AND DIFFERENTIAL PSYCHOLOGY, UNIVERSITY OF LEUVEN.

PAUL DE BOECK, PH. D., DEPARTEMENT OF DEVELOPMENTAL AND DIFFERENTIAL PSYCHOLOGY, UNIVERSITY OF LEUVEN.

ARNOLD BÖHRER, COMMANDANT, DIRECTOR BELGIAN ARMED FORCES PSYCHOLOGICAL RESEARCH SECTION.

There have been spent several efforts in showing that the low validity of personality inventories can be raised.
The approaches to this problem have been inspired by one of the followig assumptions :

(1) The validity depends on the trait that is measured.
(2) The validity depends on the criterion that is used.
(3) The validity depends on the individual that is assessed.
(4) The validity depends on the situation in which the inventory is responded to.

With regard to assumption four Claeys et al (1981) found that the validity of a personality inventory was much higher when preceded by a free selfdescription. The free-personality description used in this research inspired a new method for collecting situation-related individual personality data.

HOW TO IMPROVE THE VALIDITY OF PERSONALITY QUESTIONNAIRES

FACTORS IN THE VALIDITY OF PERSONALITY INVENTORIES

W. CLAEYS, P. DE BOECK, A. BÖHRER

Mischel (1968) has gathered an impressive amount of evidence showing that personality inventories have only a low validity when behavioral measures are used as a criterion. His conclusion is that the typical validity coefficient has a value between .20 and .30, the so-called "personality coefficient" value.

Apart from criticism on the way Mischel arrives at his conclusion (see Block, 1977), there have been spent several efforts in showing that the low validity can be raised. The approaches to this problem have been inspired by one of the following assumptions :

(1) The validity depends on the trait that is measured.
(2) The validity depends on the criterion that is used.
(3) The validity depends on the individual that is assessed.
(4) The validity depends on the situation in which the inventory is responded to.

Differential validity as a function of the trait

Mc Gowan and Gormly (1976) have shown that Mischel's conclusion may not be generalized to all of the personality traits. Dynamism, for example, like it is expressed in fysical activity, seems to be stable across situations so that individual differences on that trait may be measured in a valid way. From this finding it might be concluded that there are at least some traits lending themselves to a valid measurement. The idea of a trait difference in assessibility, due to a difference in behavioral expression intercorrelations, has been put forward already by Loevinger (1957).

[illegible] similar conclusion may be derived [illegible] work [illegible] Allen (1974) that is based on the [illegible] Allport (19[illegible]). Allport states that each individual [illegible] characterized [illegible] his own traits and that it does [illegible] other traits.

In conformity with that theory, Bem and Allen have found that not all of the individuals may be assessed on friedliness and conscientiousness in a valid way. The validity clearly depends on the cross-situational variablity of the individual within the trait domain under consideration. This finding has been replicated by Kenrick and Stringfied (1980). In the same article it is shown that also the observability of a trait has implications with respect to the validity, at least when ratings by others are used as a criterion. Traits with high observability yield clearly higher validities than traits with low observability.

Differential validity as a function of the criterion

The criteria that are used by Mischel (1968) to evaluate the validity of personality inventories are ratings or observations of behavior in specific situations. One might argue that such a criterion is too specific and not reliable enough. That is why Fishbein and Azjen (1974), Jaccard (1974), and Epstein (1980) have used combined criteria, i.e. scores summed or averaged over several situations, several times, and/or several modes of behavior expression in the same trait domain. All of these authors have shown that the validity can be raised substantially by using a combined criterion. Even validity coefficients amounting to .80 are reached.

Differential validity as a function of the individual

If it is true that individuals differ as to the degree of variability across situations, then it might be expected that traits may be measured in a valid way in some individuals (low variability individuals), but not so in others(high variability individuals). A personal style with implications to behavioral variability is the degree of self-monitoring. A high self-monitoring person has the ability and the habit of adapting himself/herself to each situation, and hence to behave in a way that is most appropriate to the given situation. His/her behavior is supposed to depend more on the situation than on personality traits. This hypothesis has been confirmed by Snyder (1974), and by Snyder and Monson (1975).

The logical consequence of this finding is that inventories should be less valid for self-monitoring individuals. That is what indeed has been found by Snyder and Swann (1976). In a study of Lippa (1978), however, this finding is only replicated with respect to neuroticism, but not to extraversion. High self-monitoring subjects are able to monitor their behavior so that they do not seem neurotic if it is appropriate, whereas low self-monitoring subjects cannot hide a possible neuroticism, even if suchlike behavior is really inappropriate in the situation in question.

Behavioral variability may be considered as an expression of the flexibility of well-adapted person in contrast to the rigidity of a maladjusted person. This consideration makes sense from studies of Moos (1968), Kogan and Wallach (1967), and Snyder and Monson (1975), so that it may be expected that personality inventories are more valid in abnormal than in normal populations.

Finally, there might be individual differences in the degree of awareness individuals have of their behavior. Individuals who are aware of their behavior may be expected to give more valid responses to a personality inventory. Awareness of own behaviors is a special case of a more general style variable that is called "self-focused attention" ("private self-consciousness"). It has been shown by Scheier, Buss, and Buss (1976) that an agressiveness inventory is clearly more valid in a group of strongly self-focused individuals (r= .66) than in a group of only weakly self-focused individuals (r= .09).

Differential validity as a function of the situation

From the research of Duval and Wicklund (1972), it may be concluded that self-focused attention can be prompted by the situation, for example by putting a mirror in front of the subject so that he/she can see him/herself. Other methods are : playing a band-recorder on which the subject hears his/her own voice, and using a TV-camera. It turned out that the self-report validity is raised under conditions favoralble to self-focused attention (Pryor, Gibbons, Wicklund, and Hodd, 1977).

An other situation that has been shown to favor inventory validity is the preceding of a free-response method of self-report (Claeys, De Boeck, and Van Den Bosch, 1981). It might be interesting to extend somewhat on this research because the results are clear but not yet published.

A first study was set up to compare the validity of a free-response method, and a structured personality inventory. The inventory (Five Personality Factors Test, 5PFT; Elshout and Akkerman, 1975) consists of five factorial scales : Extraversion, Friendliness, Conscientiousness, Neurotiscism, and General Culture. The inventory is based on the personality sphere described by Catell (1957). In the free-response method, the subjects were required to give 10 adjectives that are descriptive of their personality. After on, all of the adjectives were judged on each of the five traits of the inventory by a group of 10 experts, so that each adjective could be given a value on each trait. The averaged value of the adjectives from a subject on a trait served as a score of that subject on the trait in question. Two orders of presentation were used for the self-report : inventory/free-response method, and free-response method/inventory. The criteria to assess the validity consisted of behavioral ratings with respect to the traits of the inventory. Each subject was rated by his father, his mother and a friend. Subjects were 84 male students from the final year in a high school in a Dutch speaking town in Belgium.

The validity of the two self-report methods, the inventory and the free-response method, was about equal, but very low (r= .20), in conformity with the personality coefficient cited by Mischel (1968). A more interesting finding, however, was that the validity of the inventory was much higher (about .50 on the average (1)) when the free-description method preceded

(1) The validity coefficients that are mentioned concern a combined criterion, i.e. the summed ratings from father, mother and a friend. However, the task order effect on validity did not depend on the rater. Similar effects were found for each of the raters.

the inventory, then when the reverse order was used, in which case the validity is virtually non-existent. The same effect shows up with respect to the validity of the free-response method, i.e. its validity is much higher when it comes first than when it is preceded by the inventory.

To test the replicability of these findings, a second study was set up with essentially the same procedure and with a sample from the same population. Again the validity difference was found as a function of the task order. The validity amounts to .43 when the subject starts with the free-response method, whereas a value of only .15 is reached when the inventory precedes. A possible reason for this effect is that the free-reponse method stimulates self-focused attention much more than the personality inventory.

Discussion and Conclusion

In general, there seems to exist means to raise the validity of personality inventories. The first solution implies that only some traits may be measured. Therefore, this solution is only satisfying if one is interested in one of those exceptional traits. The second solution asks for combined criteria, whereas in practice one is mostly interested in simple criteria. The third solution restricts the use of inventories to subgroups of subjects, so that also that solution is rather unsatisfying. The fourth solution, namely providing validity favoring situations, is more hopeful, as it might be general in its applicability.

One may wonder, however, whether a concentration on the use of inventories is still appropriate in the light of a view on personality that has become less trait-based, less nomothetic, and more oriented to the individual person and his interaction with the situation, and to the processes involved in that interaction (Lamiell, 1981; Magnusson and Enler, 1977; Mischel 1973).

Such a method was developed by L.Pervin (1976,1978). Pervin asks the person to tell a number of rather important situations or events he experienced during the past year. After that, the person has to say how he percieved each situation or event, what his feelings were and how he did behave. The obtained information is put in two matrices. The first one contains situations x situational characteristics (feelings, perceptions), the second one situations x behavioral characteristics. The task of the subject is now to indicate to what degree each characteristic applies to each situation. The obtained data can be analysed by a number of different methods.

Inspired by the Pervin-method and by the free-response task used by Claeys, Böhrer tries to develop an individual personality assessment technique. The subjects are required to give 10 adjectives that are descriptive of their personality. After that, they have to indicate in which kind of situation and/or in relation with which person each given adjective is clearly appropriate and to illustrate this by a concrete event. Finaly it is asked how they feel about being (or reacting) that way. We hope, by this technique, not only to improve the validity of our measurements but also to render the selection procedure more acceptable for all subjects.

References :

Allport, G.W. Personality. A psychological interpretation. N.Y. : Holt, 1939.

Bem, D.J., & Allen, A. On predicting some of the people some of the times. The search for cross-situational consistencies in behavior. Psychological Review, 1974, 81, 506-520

Block, J. Advancing the science of personality : Paradigmatic shift or improving the quality of research ? In D. Magnusson, & N.E. Endler (Eds.), Psychology at the cross-roads: Current issues in interactional psychology. Hillsdale N.J. : Erlbaum, 1977.

Cattell, R.B. Personality and motivation structure and measurement. N.Y. : World Book, 1957.

Claeys W., De Boeck P., & Van Den Bosch W. A free-response method as an alternative to the tradional personality inventory. Unpublished manuscript, 1981.

Duval, S., & Wicklund, R.A. A theory of objective self-awaren
N.Y. : Academic Press, 1972.

Elshout, J.J. & Akkerman, A.E. Vijf Persoonlijkheidsfactoren Test (5PFT). Handleiding (Five Personality Factor Test, manual). Nijmegen : Berkhout B.V. 1975.

Epstein, S. The stability of behavior. II. Implications for psychological research. American Psychologist, 19

Fishbein, M. & Azjen, I. Attitudes toward objects as predicto of single and multiple behavioral criteria. Psycho logical Review, 1974, 81, 59-74.

Jaccard, J.J. Predicting social behavior from personality tra Journal of Research in Personality, 1974, 7, 338-

Kagan, N. & Wallach, H.A. Risk-taking as a function of the situation, the person and the group. In : G. Mandl et. al. (Eds.), New directions in psychology N° 3. N.Y. : Rinehart & Winston, 1967.

Kenrick, D.T. & Stringfield, D.D. Personality traits and the e of the beholder : crossing some traditional philosophical boundaries in the search for consistency all of the people. Psychological Review, 1980, 87, 88

Lamiell, J.T. Toward an idiothetic psychology of personality. American Psychologist, 1981, 36, 276-289.

Lippa, R. Expressive control, expressive consistency and the correspondence between expressive behavior and perso ty. Journal of Personality, 1978, 46, 438-461.

Loevinger, J. Objective tests as instruments of psychological theory. Psychological Reports, 1957, 3, 635-694

Magnusson, D. & Endler, N.S. (Eds.), Personality at the crossroads : Current issues in interactional psychology. Hillsdale, N.J. : Erlbaum, 1977.

McGowan, J. & Gormly, J. Validation of personality traits : a multicriteria approach, Journal of Personality a Social Psychology, 1976, 34, 791-795.

Mischel, W. _Personality and assessment_. N.Y. : Wiley, 1968.

Mischel, W. Toward a cognitive reconceptualization of personality. _Psychological Review_, 1973, _80_, 252-283.

Moos, R.H. Situational analysis of a therapeutic community milieu. _Journal of Abnormal Psychology_, 1968, _73_, 49-61

Pervin, L. A. A free-response description approach to the analysis of person-situation interaction. _Journal of Personality and Social Psychology_, 1976, _34_, 465-474.

Pervin, L. A. _Current controversions. Issues in personality_. N.Y. : Wiley, 1978.

Pryor, J.B., Gibbons, F.X., Wicklund, R.A., & Hodd, R. Self-focused attention and self-report validity. _Journal of Personality_, 1977, _45_, 513-527.

Scheier, M.F., Buss, A.H. & Buss, O.M. Self-consciousness, self-report of agressiveness and agression. _Journal of Research in Personality_, 1978, _12_, 133-140.

Snyder, M. Self-monitoring of expressive behavior. _Journal of Personality and Social Psychology_, 1974, _30_, 526-537

Snyder, M. & Monson, T.C. Persons, situations and the control of social behavior. _Journal of Personality and Social Psychology_, 1975, _32_, 637-644.

AD P001303

The Identification of Common Work Duties in 13 Federal Technical Occupations

Stephen M. Cormier
Personnel Research Psychologist
Office of Personnel Management

A comprehensive job analysis was conducted as the first stage in the development of a new examination for 13 occupations currently using the Junior Federal Assistant examination. These are technical occupations with the entry level at the GS-4 grade level. The need to survey a number of occupations at the same time creates potential advantages and disadvantages for the conduct of job analyses. Some of the critical issues dealt with in the course of the current project involved the level of generality of the individual task statements, incorporation of tasks from different occupations in a single inventory, and procedures for determining commonality between the different occupations. Approximately 3000 employees were selected to complete the inventory, using a sampling procedure which took into account the relative proportions of workers in the 13 occupations and the agencies and locations where they worked.

PREVIOUS PAGE IS BLANK

Although job analyses are playing an increasingly important role in personnel psychology, it still remains one of the more onerous and expensive activities in the field. In part, the expense of job analysis has been due to the requirements of the Uniform Guidelines, which set detailed standards for the job relatedness of various psychological instruments such as examinations, performance appraisal systems, training programs etc. In addition, there has been increased concern about the validation of such instruments against criteria which are reflective of actual job performance. This has also led to more elaborate job analysis techniques. Thus, there is an incentive to develop procedures which can minimize inefficiency while still retaining necessary methodological and legal characteristics.

The Junior Federal Assistant Examination Development Project, under discussion here, involves a number of factors with a bearing on the issue of the cost-effective conduct of the job analysis. These factors are: 1) the purpose for which the job analysis results will be used,2) the level of specificity of the task statements, and 3) the number of occupations to be studied. As we shall see, the first two points should be considered to be interrelated. The JFA project is designed to produce one or more selection devices that can replace or supplement the existing JFA examination. Thus, the present paper particularly applies to job analyses conducted as the initial step in examination development more than for training and classification.

Actually, there has been a general distinction between job analyses to be used for training/classification and those used as the basis for developing selection instruments (e.g. McCormick,1976,1979). The differences between these two approaches have generally been more in terms of the questions asked about the tasks than about differences in the level of specificity of the task statements. For example, a job analysis conducted for training programs might ask the raters to rate each task in terms of the time to train, the feasibility of its being trained on-the-job, the extent of training in school etc. A job analysis intended for selection purposes, on the other hand, might only ask for the criticality and time spent performing the task. However, the task statements themselves might well be identical in both cases.

Fruchter, Marin, and Archer (1963) in fact state as a theoretical ideal that all task statements should be equal in specificity, although acknowledging that practical considerations might weigh in favor of variable levels of specificity. In line with this view is the suggestion by Melching and Borcher(1973) that occupational inventories should include at least 200 but no more than 600 tasks. They argue that to include fewer than 200 tasks in an occupational area would result in task statements so general that they would yield little specific information about jobs.

Recent research is changing this picture, however, and indicating that there need be less concern about highly specific and detailed task lists in examination development. Schmidt,Hunter, and Pearlman (1981) showed in two studies comprising a massive sample size of nearly 400,000 subjects that the moderating effects of task differences had only a small impact on test validities. This proved true even when jobs were vastly different in their task composition. The validity of different test types such as verbal ability,perceptual speed,quantitative ability etc. were assessed across five distinct clerical families of jobs(stenography-typing,computing-accounting,production and stock, information and message distribution, and public contact and investigations). A comparison of the mean observed validity across these clerical job families with the pooled validity showed that there was a high degree of similarity between the test validities in different job families. In addition, the standard deviation of validities within families was not significantly smaller than the pooled standard deviation across job families. These results held true for validities using either proficiency or training success criteria.

In a second study, differences in validity were considered across Army occupations differing very substantially in their task composition (e.g. radio repair,welder,cook,dental assistant). The same kind of comparisons between mean and pooled validities were conducted for the basic test types used in the first study. A high correlation was found between the validity estimates of the first study with those obtained in the second. The true standard deviation for the test type validities was found to be .1081. As Schmidt et al. note, for SDs of this size,very large sample sizes are needed to detect the moderating influence of different jobs, even when the difference is one or two SDs. Since the average validity of these different kinds of aptitude tests was .45, 90% of the validities across different jobs have values between .27 and .63. Even at the lower end this is a respectable and useful validity. In addition, it should be remembered that the variation among jobs is about as extreme as will ever be found in normal organizational situations.

Other studies have also reported substantially similar validities across occupations (e.g Ghiselli,1966; Maier & Fuchs,1969). For example, Ghiselli found that clustering jobs according to similarities of average validity patterns of aptitude tests resulted in the grouping of jobs which had little apparent similarity in the nature of their respective tasks. He concluded that jobs appearing to be different in terms of the nature of the work may require similar job-related abilities (cf. Mobley & Ramsay,1973; Randhawa,1978).

Development of the Inventory

The occupations included in the present job analysis were those series identified as primary users of the current JFA exam for selection. The exam is being used for entry into a number of technical occupations in the Federal government at the GS-4 level. Although certain other occupations make selections off the JFA register, this is done out of convenience rather thas as a result of occupational standards. The 13 occupations covered in the present study are listed in Table 1.

It is obvious that the occupations using the current JFA exam are a diverse group. The central question of the job analysis is the degree of underlying similarity [illegible] in the [illegible] of these different series. In addition to the validity generalization issue raised previously, there are two other issues that should be considered. First, although two series might be dealing with different subject matter, the tasks assigned to the technicians might include essentially the same activities. Secondly, it is entry-level positions that are going to be filled through whatever selection devices are developed. While specialized knowledge may be important at higher grade levels, such knowledge is not required at the entry-level.

As noted, one of the major concerns of this paper is the question of efficiency in the conduct of job analyses. This is particularly important when a number of different occupations are being surveyed at the same time. At the outset, a number of basic decisions were made on the development of an appropriate task inventory for the 13 occupations based on the research literature and the purpose of the study. First of all, given the accumulating evidence on the small impact of task differences on test validities, it was deemed unnecessary to develop long lists of molecular tasks for each of the 13 occupations. Instead, it was felt that the level of specificity of the statements should be somewhere between an overall duty and a particular task. Secondly, since the purpose of the analysis was to determine occupational commonalities, it made sense to use a single inventory containing the duty/task statements for all the occupations for three reasons: 1) the reduction in the number of task statements for each occupation made it feasible to include them all in the same booklet, 2) all the respondents could be exposed to the same set of tasks to rate, thereby making the measurement of response commonality possible, and 3) in some locations, different occupations can be found together and surveying bercomes easier when the inventory can be used by all occupations.

The development of initial lists of tasks was started in the usual way by examining descriptions and standards for the occupations. Once an initial list of tasks for each occupation was extracted from these materials, they were put into a form which would meet the requirements concerning comprehensibility, observability etc,. mentioned in the Uniform Guidelines. Besides these more or less technical revisions, the general question of task specificity was examined while rewriting the duty/task statements.

The main concern was to have the duty/task statements preserve fundamental and real differences in the nature of tasks without reflecting minute or terminological differences. Unifying factors existed since the overall grade levels under consideration for the occupations are equivalent and thus the level of difficulty should be as well. With respect to task differences, two occupations may have tasks involving the filing of different forms; however, the content and level of difficulty could be essentially the same. It was decided to try to rewite task statements so that differences in substance rather than form would be preserved. Subject-matter experts could provide an independent source of judgement when difficulties of interpretation arose. In this way, certain tasks could be collapsed across occupations into a more general duty/task. Once the rough form of the inventory was determined, panels of SMEs from each occupation evaluated the task lists for their jobs. After the review panel for each occupation had gone over the inventories and agreed that the lists were accurate and comprehensive,the inventories were changed to reflect their comments.

The final inventory was composed of a background section containing a number of biographical questions to ensure that a representative sample had been obtained. Following this was the list of 222 tasks for all 13 occupations. The tasks were arranged in alphabetical order rather than by occupation to ensure that the respondents would look at all the task statements. Each duty/task was to be rated on performance,time spent, and importance scales.

Administration of the JFA Inventory

In order to ensure that the results of the job analysis were representative of the work force, a sampling plan was developed. Essentially, a probability sample was created using occupation,agency, and location as the controlling factors. Each occupation has a different population size than the others and the total range is quite large (84 employees in series 992 compared with 25,960 in series 2005). Due to the wide variation in occupational size, it was impossible to maintain the exact proportions in the sample, although an effort was made to approximate the correct proportions.

Other important considerations that were desirable to take into account were the age,ethnic composition,grade level and sex of the respondents in the sample. These last variables were more subject to the individual agencies than to OPM.

Approximately 3000 employees were sampled in the 13 occupations being analyzed. The respondents worked in various locations around the continental United States. Only those employees in grades 4-7 were sampled since the Guidelines indicate that only those tasks likely to be asked of the average employee within 5 years of entry into the occupation should be considered in entry level exam development.

Results and Discussion

Of the 2997 inventories that were completed, 211 had to be discarded because they were too incomplete to process (e.g. the importance scale was not filled out), respondents were not in grades 4-7, or key biographical data was miscoded. Since the original probability sample had specified about 2000 respondents some deletion of inventories was necessary. This was particularly the case since the overrepresentation was not consistent from occupation to occupation. A computer program was used to randomly select the correct proportions from each occupation.

Using the biographical data, it was possible to ascertain that the sample matched the characteristics of the overall population on almost all of the relevant variables, including minority representation.

Since,as expected, there was a high degree of correlation between the ratings on time spent and importance on the duty/tasks,it was only necessary to analyze the data on the latter scale. The duty/task ratings on importance were first analyzed using the method advocated by Christal(1974). Each task rating was converted to a percentage of all the ratings from each respondent. Then these proportions were averaged over all raters and not just those performing the task.

The importance ratings for all the duty/tasks were analyzed at both the occupational and combined level. A rational criterion was set in order to identify the most important tasks. The criterion was suggested by the form of the data: there seemed to be a natural dichotomy between the top 10 or 15% of the rat'd duty/tasks and the remainder, at both levels of analysis. In general, duty/tasks rated by at least 25-30% of the respondents had a good chance of being included in the top group.

At the overall level,32 of the 222 duty/tasks exceeded the criterion mean . Approximately the same number of duty/tasks exceeded the criterion at the occupational level as well. Although each occupation had critical tasks idiosyncratic to the occupation, there was a great deal of overlap between the critical tasks at the occupational level and at the combined level. Table 2 lists the percentage overlap between the critical duty/tasks of each occupation with the critical duty/tasks identified at the combined level. As can be seen, all but one of the occupations has an overlap of approximately 50% or more with the overall critical duty/tasks. The exception is the 1105 series with a 35% overlap. At the upper end we have the 1107 series with a 90% overlap.

In order to obtain a more precise view of the interrelationship among the 13 occupations, cluster analyses were performed on the importance ratings from the various occupations. A method proposed by Johnson (1967) was used because of its compatibility with data possessing only known rank-order properties. The algorithm contains the following properties: 1) input consists solely of $n(n-1)/2$ similarity measures among the $\underline{n}$ objects under study and 2) the clustering is essentially invariant under monotone transformations of the input similarity data.

The similarity matrix was constructed using the measure described by Archer (1966). The overlap between any pair of $\underline{p}$ variates in the form of relative frequencies is found by selecting the minimum proportion from each pair and then forming the sum of such pairs. A sum of 0 indicates no correspondence while a sum of 1 indicates a perfect correspondence.

Johnson specifies two possible approaches to the identification of cluster or similarity solutions, namely, the minimum and maximum methods. In the maximum method, for each clustering the diameter of the cluster is computed. The value of the clustering is the maximum diameter of the various clusters. The minimum method follows the same initial steps as the maximum procedure; however, the distance between any sequence of objects is determined. The procedure can identify long serpentine-like shapes. The total size of a chain is the largest link distance, while the chain distance between any two clusters is the minimal chain size of all chains between the two.

In the present study both procedures were used although it was felt that the the minimum or connectedness method would probably yield the most interpretable results. This is because of the inherent nature of the task data across occupations; each occupation will have a number of idiosyncratic tasks that will make compact single clusters more unlikely but linkages more likely. A similarity matrix was computed for the importance data from the 13 occupations. This similarity matrix was then analyzed according to the minimum and maximum methods with the results appearing in Tables 3 and 4 respectively. The graphical display is generally in the form of a right triangle where objects are merged one by one until all are included. As expected the minimum method produced somewhat more coherent results although there was significant inter-agreement. In Table 3 it can be seen that there are two main clusters of occupations; the first comprises series 962,992,1105,1107, 1411,and 1421, while the second comprises series 963,986,990,1702, and 2005. Even these two clusters quickly become merged at a relatively high level(.599). The 344 and 1106 series on the other hand seem to be anomalous in terms of their response pattern since the incusion of these two series substantially decreases the index of similarity (.250).

In Table 4 we see the results from the maximum or diameter method. Again, series 344 and 1106 appear to be the most resistant to clustering. The two main clusters of the connectedness method are divided into four,possibly for the reasons discussed above. The various clusters do merge, but at a lower level than in the connectedness method (.20). It should be noted in connection with the two anomalous series that there is some indication that duties assigned by the employing agencies have a significant amount of heterogeneity in these series. This may account for their position as outliers.

The results indicate that at least 11 of the 13 occupations possess substantial commonality among themselves, based on the overall pattern of importance ratings on the duty/tasks. Since even with the two outlying series there was substantial overlap on the most critical tasks, the commonality may even be stronger than suggested by the cluster analysis, at least as far as the Guidelines are concerned. The Guidelines indicate that different jobs can be grouped together for validity studies when they have substantially the same major work behaviors(U.S. EEOC et al. p.38300,section 14b) .

Similar findings for clerical occupations have been reported by Abbe(1980). A single large cluster of clerical occupations were established using a variety of techniques, including the present approach,factor analysis, and principal component analysis. Thus, there is good reason to believe that more cost-effective methods of job analysis can be employed in examination development consistent with legal and methodological requirements.

Table 1

The 13 Occupations Used in the Study

1. GS-344 Management Assistant
2. GS-962 Contact Representative
3. GS-963 Legal Instruments Examiner
4. GS-986 Legal Technician
5. GS-990 General Claims Examiner
6. GS-992 Loss and Damage Claims Examiner
7. GS-1105 Purchasing Agent
8. GS-1106 Procurement Technician
9. GS-1107 Property Disposal Technician
10. GS-1411 Library Technician
11. GS-1421 Archives Technician
12. GS-1702 Education and Training Technician
13. GS-2005 Supply Technician

Table 2

Proportion of Overlap Between the Most Important Tasks at the Combined Level with the Occupational Level

Series	Proportion of Overlap
344	.65
962	.56
963	.55
986	.73
990	.48
992	.47
1105	.35
1106	.74
1107	.90
1411	.50
1421	.62
1702	.81
2005	.81

Table 3

Cluster analysis using the connectedness method

Similarity Index	Occupations 1.		8.		10.		6.		7.		2.		9.		11.		4.		12.		5.		3.		13.
.99200500E 05	.		.		.		.		.		.		.		.		.		.		.		X	X	X
.98800400E 05	.		.		.		.		,		X	X	X		.		.		.		.		X	X	X
.96700500E 05	.		.		.		.		.		X	X	X	X	X		.		.		.		X	X	X
.94500600E 05	.		.		.		.		.		X	X	X	X	X		X	X	X		.		X	X	X
.82400400E 05	.		.		.		.		X	X	X	X	X	X	X		X	X	X		.		X	X	X
.76100500E 05	.		.		.		X	X	X	X	X	X	X	X	X		X	X	X		.		X	X	X
.82400300E 05	.		.		X	X	X	X	X	X	X	X	X	X	X		X	X	X		.		X	X	X
.75100700E 05	.		.		X	X	X	X	X	X	X	X	X	X	X		X	X	X		X	X	X	X	X
.76100500E 05	.		.		X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X
.59900400E 05	.		.		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
.26000700E 05	.		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
.25000700E 05	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Table 4

Cluster analysis using the diameter method

Similarity Index	Occupations 5.		7.		1.		8.		4.		12.		6.		10.		2.		9.		11.		3.		13.
.99200500E 05	.		.		.		.		.		.		.		.		.		.		.		X	X	X
.98800400E 05	.		.		,		.		.		.		.		.		X	X	X		.		X	X	X
.94500600E 05	.		.		.		.		X	X	X		.		.		X	X	X		.		X	X	X
.89800500E 05	.		.		.		.		X	X	X		.		.		X	X	X	X	X		X	X	X
.78900500E 05	.		.		.		.		X	X	X		.		.		X	X	X	X	X	X	X	X	X
.74500300E 05	.		.		.		.		X	X	X		X	X	X		X	X	X	X	X	X	X	X	X
.70800400E 05	X	X	X		.		.		X	X	X		X	X	X		X	X	X	X	X	X	X	X	X
.26700600E 05	X	X	X		.		.		X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X
.20200400E 05	X	X	X		.		.		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
.15000700E 05	X	X	X		X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
.30000500E 04	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
.40000000E 00	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

References

Abbe,E.N. *A methodological strategy for identifying similarities among jobs.* (PRR 80-25) Washington D.C.: Office of Personnel Management,1980.

Archer,W.B. *Computation of group job descriptions for occupational survey data* (PRL-TR-66-12). Lackland AFB,Texas: Aerospace Medical Division,1966.

Fruchter,B., Marin,R.E.,& Archer,W.B. *Efficiency of the open-ended inventory in eliciting task statements from job incumbents.* U.S. Air Force,AMD/AFSC,Personnel Research Laboratory,Lackland AFB,Texas,1963

Ghiselli,E.E. *The validity of occupational aptitude tests.* N.Y.: Wiley,1966.

Johnson,S.C. Hierarchical clustering schemes. Psychometrika,1967, 32,241-254.

Maier,M.H. & Fuchs,E.F. *Development of improved aptitude area composites for enlisted classification.* Technical Research Report 1159. Arl.,Va.:U.S. Army Behavioral Science Research ,1969.

Melching,W.H. & Borcher,S.D. *Procedures for constructing and using task inventories.* Center for Vocational and Technical Education, Research and Development Series v.91. Cloumbus,Ohio: Ohio State U. 1973.

Mobley,W.H. & Ramsay,R.S. Hierarchical clustering on the basis of inter-job similarity as a tool in validity generalization. *Personnel Psychology*,1973,26,213-25.

Morsh,J.E. & Archer,W.B. *Procedural guide for conducting occupational surveys in the U.S. Air Force.*(PRL-TR-67-11) . Lackland AFB Texas: Aerospace Medical Division,1967.

Randhawa,B.S. Clustering of skills and occupations: A generic skills approach to occupational training. *Journal of Vocational Behavior*, 1978,12,80-92.

Schmidt,F.L.,Hunter,J.E. & Pearlman,K. Task differences as moderators of aptitude test validity in selection: A red herring. *Journal of Applied Psychology*,1981,66,166-85.

U.S. EEOC et al. Uniform guidelines on employee selection procedures. *Federal Register*,Aug.25,1978,v.43,no.1666,38295-38309.

Ward,J.H. & Hook,M.E. Application of an hierarchical grouping procedure to a problem of grouping profiles. *Educational and Psychological Measurement*,1963,23,69-81.

COMPUTERIZED ADAPTIVE TESTING: FROM PSYCHOMETRICS TO SYSTEMS

Paul R. Croll
Director, CAT System Development Project
U.S. Office of Personnel Management

The design of a Computerized Adaptive Testing (CAT) system requires the careful integration of psychometric and engineering developments. System designers must be cognizant of the relationships among the psychometric components of the system, among the physical components of the system, and between its psychometric and physical components. A hierarchical functional design model has been developed to facilitate the design of a prototype CAT system for DoD enlisted personnel accessions testing. The model addresses both the psychometric and administrative/operational requirements of the system and serves as the blueprint for translating these functional requirements into a working prototype. The focus on function in this design approach insures that the description of CAT psychometric procedures, in terms of system functions, precedes specification of the physical elements of the system through which these procedures are implemented. Such precedence is critical to the psychometric integrity of the system development effort.

Computerized adaptive testing (CAT) is a remarkably effective combination of recent developments in latent trait theory and continuing advances in computer technology (Urry, 1977a). Unlike conventional paper-and-pencil group testing, in which identical test forms are administered simultaneously to large groups of examinees, computerized adaptive testing is an individualized testing procedure that constructs, administers, and scores tests interactively during the actual testing session. Each examinee receives only those questions appropriate to his or her own level of ability, resulting in an individualized test "adapted" or "tailored" to the specific examinee's level. The number of questions required to produce an estimate of ability at the same level of reliability as the longer group test is considerably less. In 1979, the Department of Defense established a joint-service project, led by the Navy Personnel Research and Development Center, to evaluate the feasibility of CAT for enlisted personnel accession testing. The project has been conceived as a large-scale system development effort, integrating psychometric and engineering developments to meet system goals.

CAT System Design Principles

The primary objectives of the CAT system development effort are the design, development, test and evaluation of a system for automated adaptive administration of the Department of Defense's enlisted personnel selection and classification tests. The desired outcome of the development effort is an integrated set of well-defined inputs, processes, and products meeting the following criteria: (1) user (i.e. service) needs are readily translated into specifications that both define system products and provide for control of system processes; (2) system products completely and consistently conform to user specifications; and (3) system processes and products are continually monitored to insure such conformance. The capability for delivery of well-defined products, meeting user needs and monitored for conformance with user specifications, is the essence of a CAT *system*.

The system development problem is approached through two distinct lines of development: psychometric development of the procedures for adaptive testing and engineering development of the physical system through which these procedures will be implemented. The application of system design principles to the development of the computer-based physical system is straightforward and well supported by present practice. The application of such principles to the development of psychometric procedures, however, is unique and can present a subtle danger to the integrity of the system as a whole. The danger lies in the possible failure to recognize that the CAT system must be designed to meet psychometric objectives first. Engineering objectives must not be permitted to drive the system development effort. For example, modification of well-proven CAT algorithms, to fit an initial conception of hardware performance characteristics, is inappropriate. Rather, the algorithms chosen should dictate hardware specifications. Viewing CAT system development as simply another data processing system development exercise is likely to compromise the system's psychometric integrity. Recognition of the tremendously complex network of interactions underlying systems design is especially necessary in CAT system development. System designers must be cognizant of the relationships among the psychometric components of the system, among the physical components of the system, and between its psychometric and physical components. Appreciation of these relationships is critical in integrating these components into a well-functioning system.

In order to facilitate such integration, the design strategy chosen for the CAT system has focused on function rather than structure. A hierarchical functional design model has been developed addressing both the psychometric and administrative/operational requirements of the system. This model consists of a set of hierarchical functional descriptions of system components and their interrelationships. These detailed descriptions serve as the basis for design of the system structure, system prototyping, and final system development. A useful technique for developing such functional descriptions is known as HIPO (Hierarchy plus Input-Process-Output) (IBM, 1975; Katzan, 1976). HIPO is a technique that describes system functions in terms of inputs, processes, and outputs. These functional descriptions are presented hierarchically, showing in greater and greater level of detail the functional relationships within and between system components. Use of the HIPO technique aids system development, in that all required inputs, processes, and outputs at each level of functional detail are specified.

Katzan (1976) describes a system function as a process that accepts one or more inputs and produces one or more outputs. The application of this definition in computer hardware or software design is straightforward. For example, the *multiply* function of a CPU chip accepts a multiplier and multiplicand, each of fixed length, and returns a product. Valid input sources and output destinations are inherent in the chip design. The application to software design is analogous, with the program code determining input sources and characteristics, output destinations and characteristics, and the intervening processing steps necessary to produce output from input. The application of this definition to the design of a psychometric system is less obvious. Even Chapanis, writing on both human factors in systems engineering and on systems staffing, in de Greene's text on systems psychology (1970a, 1970b), neglects the application of system design principles to the development of psychometric procedures. Systems thinking is only applied to the problem of personnel selection and classification, and then only in the sense that a systematic approach to selecting, evaluating, and training personnel is seen as a component of a larger system design. Systems thinking need not stop short with the human factors or engineering psychology approach, however. It is readily applicable to basic psychometric developments as well.

If one defines a personnel measurement procedure as the administration, scoring and evaluation of the results of a test of some ability, questions couched in system design terms can easily be raised. What are the desired outputs: test records, scores, selection decisions? What are the processes required to obtain those outputs: administration of test questions recording of examinee responses, scoring, application of selection rules? What are the inputs required by the specified processes to produce the desired outputs: instruction sets, test questions, examinee responses, scoring keys? To be sure, this is a simplistic example. However, it does illustrate that psychometric issues such as personnel measurement may be addressed from a system design perspective, bringing to bear all the tools and techniques of that discipline. The design of a CAT system is a far more complex undertaking, yet the development of a functional design model for the system greatly simplifies the dual tasks of psychometric and engineering development and facilitates their eventual integration.

Functional Overview of a Computerized Adaptive Testing System[1]

In computerized adaptive testing, tests are constructed, administered, and scored interactively during the actual testing session. What functions are necessary to this process? First, it is obvious that a function encompassing test construction, administration and scoring can be defined. Is this sufficient? Where do the test questions to be administered come from? In CAT, test questions for each ability to be tested are selected for administration from a set of questions called an item bank. Item banks are carefully constructed sets of test questions having well-specified psychometric properties, with each item bank designed to measure a single ability. It then becomes obvious that a function providing for item banking must also be defined. Are these functions now sufficient? Remember that in CAT, a

---

[1] The development of a functional design model for a CAT system has been based on both the analysis of NPRDC-specified requirements and objectives and the author's experience with the design of a similar system at the U.S. Office of Personnel Management (see Croll & Urry, Note 1).

test is terminated when a pre-specified level of reliability is reached. Also, in multiple ability testing, a weighted composite score may be required. Where do the termination rules and score weights come from? It seems, then, that a function providing such measurement control parameters is also required. Have all the functions necessary for CAT now been defined? In a perfect world perhaps, but in the real world things go wrong. How will we know if things go wrong in CAT? A function providing for monitoring of CAT functioning and for quality control reporting would do nicely.

Through the application of such a simple functional analysis to the CAT process, the four major functions of a computerized adaptive testing system have been identified.

A HIPO package for the initial design of a CAT system has been developed. From the visual table of contents (Figure 1) and the CAT system overview diagram (Figure 2), it can be seen that the four major system functions have been identified: item banking, measurement control, test administration and scoring, and monitoring and quality control. Outputs of the item banking and measurement control components are required as inputs to the test administration and scoring component, and outputs from the test administration and scoring component are required as inputs for monitoring and quality control.

The Item Banking Function

This component of the CAT system provides the sets of test questions, or item banks, necessary for adaptive test administration. It is composed of three subfunctions: test item calibration, item bank construction, and item bank evaluation.

Test item calibration refers to the estimation of the latent trait parameters a_i, b_i and c_i of candidate test questions for item banking (Urry, Note 2). Input for this subfunction consists of results from either conventional or adaptive administration of the potential test questions. If parameters are to be estimated from conventional test results, examinee response data and scoring keys for the questions must be supplied. If parameters are to be estimated from adaptive test results, ability scores must be supplied as well.

An algorithm for the estimation of parameters from conventional test results has been described by Urry (Note 3). Schmidt and Urry (1976) have also described the results of an algorithm for the estimation of parameters from adaptive test results. This algorithm is also described in Urry (Note 3). These algorithms are suggested as a guide for the design of the system's parameter estimation subfunctions. Parameter estimation from adaptive test results is especially important in CAT in that it permits on-line calibration of potential test questions in the normal operational context of the CAT system. It provides a means to eventually end dependence on conventional test results for item parameter estimation.

The test item calibration subfunction outputs parameter estimates and calibration statistics for the potential test questions. The parameter estimates are then treated as input to the item bank construction subfunction.

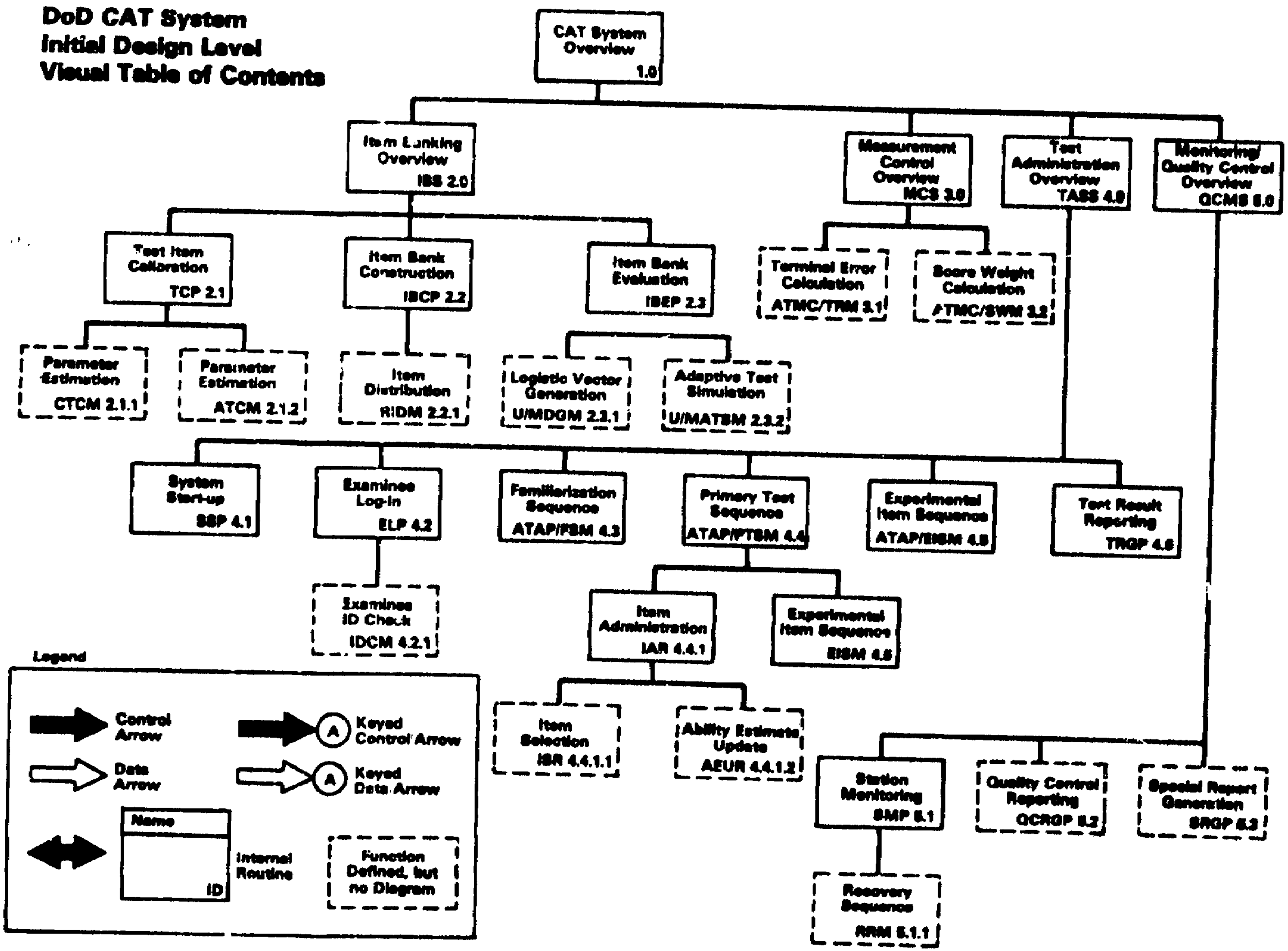


Figure 1. The visual table of contents for the DoD CAT system initial design level HIPO package.

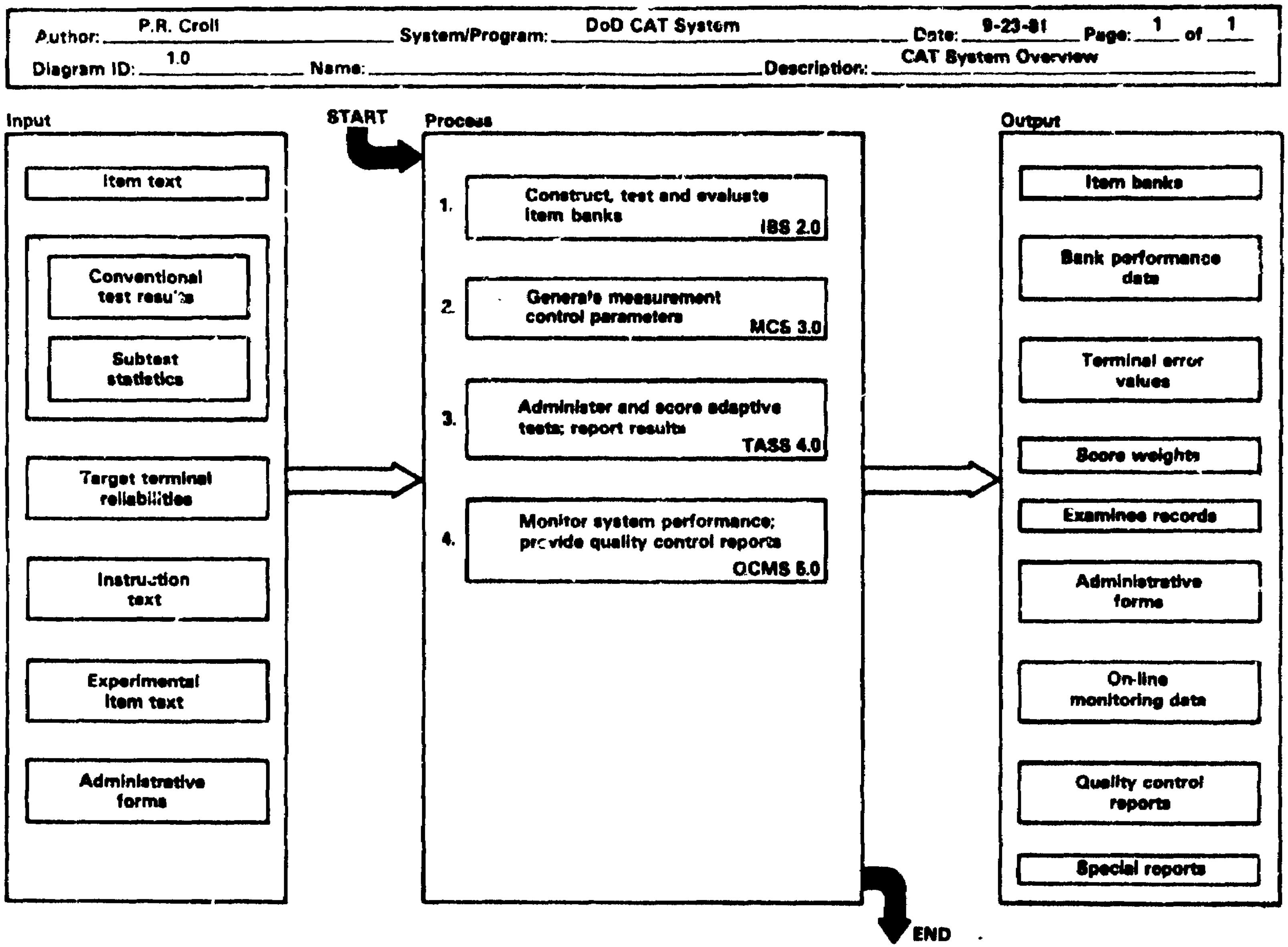


Figure 2. Functional overview of the DoD CAT system

The item bank construction subfunction takes the parameter estimates for candidate questions and compares them against target values for the a_g and c_g parameters. The prescription for acceptable values of these parameters has been detailed by Urry (1970, 1977b; Note 2). Questions failing to meet this prescription are rejected. The remaining item parameter sets are then sorted to ease later processing and a rectangular distribution of the items, by the b_g parameter, is built. Urry's prescriptions for the size and distributional shape of an item bank may then be followed in selecting questions for inclusion. The tentative item bank produced by this subfunction is then evaluated in the item bank evaluation subfunction.

The item bank evaluation subfunction is designed to assess the performance characteristics of an item bank before it is placed into operational use. It is a critical component of the CAT system design, in that item bank performance characteristics are a major determinant of CAT system performance. A procedure for the evaluation of an item bank has been described by Urry (1974). From the functional perspective, the item parameter sets for the tentative item bank are used to generate response vectors (1's and 0's, or rights and wrongs) for simulated examinees. Termination rules are selected for item bank evaluation, based on the desired reliability of measures to be constructed from the bank (Urry 1977b, Note 2). These rules are provided by setting a prespecified value of the error of the ability estimate, at which the test sequence will be terminated. Adaptive testing is then simulated using the item parameter sets, response vectors, and termination rules, and the results are reported. Only if the item bank is judged acceptable is it made available, with associated question text, for operational use. The procedural steps in the item banking function are repeated for each ability for which an item bank is to be constructed.

Additionally, when several item banks are to be administered as a multiple-ability battery, simulation of adaptive testing with the complete set of banks is conducted.

The Measurement Control Function

This function is the most critical component of the CAT system. It provides the means through which user (i.e. service) answers to the three basic questions underlying CAT system operation are translated into system control parameters. These three questions are (1) "What is to be measured?"; (2) "What degree of accuracy is to be employed?"; and (3) "How are subtest scores to be combined into composite scores?". Without such a function, users have no way of communicating their requirements to the system. User requirements are communicated to system personnel who, in turn, specify the measurement protocols required to meet the user's needs. These protocols embody the measurement requirements of each system user and determine both the way in which the adaptive testing process proceeds and the nature of outputs of that process. They specify the combination of subtests required to meet specific measurement objectives (e.g. full-ASVAB vs. AFQT, or service-specific composites), the outputs desired (e.g. subtest scores vs. weighted composite scores), the scale of measurement desired, and the accuracy of measurement desired. They take the form of the input stream required by the system to generate control parameters.

It is through software generation of control parameters that user measurement protocols are implemented in the CAT system. These parameters are of three types: (1) termination rules, or terminal error values (values for the error of the estimate of ability), that determine the point in the adaptive testing sequence at which testing for a particular ability is terminated; (2) subtest weights, that determine the relative contribution of a subtest score to a composite score (and which may be zero, if a subtest score is not to be included in a particular composite score); and (3) rescaling factors, which provide for the conversion of scores based in the system's standard scale of measurement to an alternate scale of measurement.

The Measurement Control function must provide the capability for translation of a wide range of user measurement protocols into appropriate control parameters. It can become quite complex as the number and complexity of distinct user protocols increases. The psychometric bases for this function have been discussed by Urry (Note 2, Note 3). Its implementation is dependent upon several necessary conditions of the total system design: (1) a Bayesian modal solution for item parameter estimates must be used; (2) the Owen Bayesian algorithm must serve as the basis for item selection and ability estimation; and (3) a variable test length termination strategy, based on target values of the standard error of the estimate of ability (for each subtest), must be employed.

The Test Administration and Scoring Function

This component of the CAT system provides for the administration and scoring of adaptive tests in the live testing environment. It is what is often thought of as the sole function of a CAT system, since it is the primary system function that is implemented in the field-resident physical system. It is composed of six subfunctions.

The system start-up subfunction is composed of those steps necessary to prepare the physical system (i.e., the hardware and software) for a testing session. It includes power-up, self-test, sign-on, and system status verification activities.

The examinee log-in subfunction performs the administrative tasks required to identify the examinee to the system and to link the examinee's test record with the other steps in the applicant processing sequence. Inputs include data from administrative forms and examinee supplied data. Outputs include administrative forms and the examinee record into which the test results will later be written. Additionally, a lower-level subfunction has been specified that insures that examinees are correctly seated at the testing stations to which they have been assigned.

The familiarization subfunction is designed to familiarize the examinee with both the hardware to be used in taking an adaptive test and the adaptive testing process itself. Introductory, instructional, and practice material is displayed for the examinee on the testing station display, and the examinee enters the required responses on the testing station keyboard. Checks are included to insure that the examinee is proceeding through the familiarization sequence successfully. An option has also been designed for the examinee to request a repeat of the familiarization sequence. Inputs include introductory, instructional, and practice text and examinee responses. Outputs consist of displays of the input text and error messages.

The primary test subfunction is the heart of the test administration and scoring function. It is designed to select and display test questions, read and score examinee responses, and update the examinee test record. Additionally, it provides for the administration of experimental items (through branching to another subfunction), selective retests, and test result recording on the testing site's master file. Inputs include control parameters, item parameters, item text, and examinee responses. Outputs include test item displays, error message displays, and the examinee test record.

Several lower-level subfunctions have been specified. The item administration subfunction selects and displays test questions, reads examinee responses, and displays an error message when appropriate. It also scores examinee responses and updates the estimate of ability and its associated error value. Additionally, it provides for termination of the testing sequence in a particular ability by checking the current error value of the ability estimate against the pre-specified terminal error value. Since the item selection and ability and error updating procedures are psychometrically complex, lower level subfunctions have been identified for these procedures, but not specified in separate HIPO diagrams. Decisions regarding the nature of these subfunctions will have to be made within the context of the system's psychometric development activities. The reader is referred to Urry (1977b; Note 2, Note 3) for guidance in developing these procedures.

The experimental item subfunction provides for the administration of experimental, or potential, test questions within the context of an adaptive test. It selects and displays experimental items and reads and records examinee responses. Inputs include item bank codes, item text and examinee responses. Outputs include item text displays and examinee reponses to the items. This subfunction is called by the primary test subfunction, when control codes indicate that experimental items are to be administered.

The test result reporting subfunction is designed to provide printed reports of test results, including any required administrative forms. It inputs data from the testing site's configuration master file and prints reports as required. It is also designed to feed testing results into the AFEES Reporting System.

The Monitoring and Quality Control Function

This component of the CAT system provides for system-wide quality control of all CAT system functions, as well as for monitoring of the actual testing process at the testing site. It is composed of three subfunctions: testing station monitoring, quality control report generation, and special report generation. The term "quality control", as used in this function, does not simply imply physical system diagnostics and maintenance, but also implies monitoring and control of the psychometric integrity of the CAT system. Since the system will stand or fall on the quality of its personnel measurement, its psychometric integrity requires constant scrutiny.

Some suggestions for the testing station monitoring subfunction are offerred. Three conditions might occur during a testing session which would require the attention of the test monitor: the examinee may fail to progress normally through the testing sequence, and also fail to request assistance; the examinee may, for any reason, request test monitor assistance; and a failure might occur in a testing station. Also, psychometric anomalies may occur, such as the presentation of a sequence of items that is excessively long given the characteristics of the item bank. The testing station monitoring subfunction should provide a constant display of testing station status, so that such conditions might be identified. Additionally, if a testing station fails, a lower level subfunction should be initiated to perform a recovery/restart sequence.

The four major functions identified in the CAT functional design model suggest a system structure which implements each of these functions in a separate subsystem, each with its own data, logic, hardware, and software characteristics. It is through their accurate translation into components of system structure and the implementation of system structure in the physical system, that CAT will finally have made the transition from psychometrics to systems.

References

Birnbaum, A. Part 5. Some latent trait models and their use inferring an an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, Massachusetts: Addison & Wesley, 1968.

Chapanis, A. Chapter 2. Human factors in system engineering. In K. B. de Greene (Ed.), Systems psychology. New York: McGraw-Hill Book Company, 1970(b).

Chapanis, A. Chapter 12, System Staffing. In K. B. de Greene (Ed.), Systems psychology. New York: McGraw-Hill book Company, 1970(b).

IBM Corporation. HIPO - A design aid and documentation technique, Second edition. White Plains, New York, 1975. (IBM order number GC20-1851-1,

Katzan, H. Systems design and documentation. An introduction to the HIPO method. New York: Van Nostrand Reinhold Company, 1976.

Schmidt, F. L. & Urry, V. W. Item parameterization procedures for the future. Proceedings of the First Conference on Computerized Adaptive Testing (PS-75-6, Personnel Research and Development Center, U.S. Civil Service Commission). Washington, D.C.: U.S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-00-00940-9).

Urry, V. W. A Monte Carlo investigation of logistic metal test models. test models. (Doctoral dissertation, Purdue University, 1970). Dissertation Abstracts Intenational, 1971, 31, 6319B. (University Microfilms No. 71-9475).

Urry, V. W. Computer-assisted testing: calibration and evaluation of the verbal ability bank (TS-74-3). Washington, D.C., Personnel Research and

Developement Center, U.S. Civil Service Commission, December, 1974. (NTIS No. PB 261 694/AS).

Urry, V. W. Ancillary estimators for the item parameters of mental test models. In W. A. Gorham (Chair), Computers and testing: steps toward the inevitable conquest. Symposium presented at the 83rd Annual Convention of the American Psychological Association, Chicago, 1975. (NTIS No. PB 261 694/AS).

Urry, V. W. A five-year quest: Is computerized adaptive testing feasible? Proceedings of the First Conference on Computerized Adaptive Testing (PS-75-6, Personnel Research and Development Center, U.S. Civil Service Commision). Washington, D. C.: U.S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-00-00940-9).

Urry, V. W. Tailored testing: a successful application of latent trait theory. Journal of Education Measurement, 1977(a) 14, 181-196.

Urry, V. W. Tailored testing: a spectacular success for latent trait theory. (TS-77-2). Washington, D. C.: Personnel Developement Center, U.S. Civil Service Commission, August, 1977(b) (NTIS No. PB 274 576).

Reference Notes

1. Croll, P. R., & Urry, V. W. Tailored testing: Maximizing valdity and utility for job selection. Paper presented as part of a symposium on Innovations in Personnel Testing, 86th Annual Convention of the American Psychological Association, Toronto Canada, 1975.

2. Urry, V. W. Tailored testing: Its theory and practice. Part 1. The basic model, the normal ogive submodels, and the tailored testing algorithms. San Diego, California, Navy Personnel Research and development Center. In press.

3. Urry, V. W. Tailored testing: Its theory and practice. Part 2. Ability and item parameter estimation, multiple ability applications, and allied procedures. San Diego, California, Navy Personnel Research and Development Center. In preparation.

Acknowledgments

The work reported in this paper was performed under a reimbursable services agreement, for the Navy Personnel Research and Development Center, San Diego, California. The views expressed in this paper, however, are those of the author and do not necessarily reflect those of the Navy or the Department of Defense. The author wishes to thank Dr. Vern W. Urry for the knowledge, training, and experience he has provided over the last several years; and Dr. James McBride, for his support in pursuing this area of investigation.

AD P001305

DEVELOPMENT OF A WORK BEHAVIOR SURVEY FOR THIRTY-FIVE FEDERAL CLERICAL OCCUPATIONS

Cynthia C. Diane
Personnel Research Psychologist
United States Office of Personnel Management

A comprehensive work survey form was developed to study thirty-five clerical occupations in the Federal sector. The purpose of the survey was to develop new selection procedures which are not only job related, but which also meet the requirements of the Uniform Guidelines on Employee Selection Procedures.

Work behavior statements were obtained from previous job analysis studies, existing inventories and the Dictionary of Occupational Titles. The statements were then reviewed by three experienced incumbents in each of the thirty-five occupations. The final survey form incorporated the comments of these subject matter experts, and contained a listing of 174 work behaviors which crossed over the thirty-five occupations being studied. The survey form has been completed by approximately 3000 incumbents, nationwide. Data analysis will show the degree of overlap between these occupations and aid in the development of new selection procedures.

The decision by the Office of Personnel Management (OPM) to undertake a job analysis of clerical occupations was influenced by two things. First, the clerical occupational group is the most populous in the Federal sector. At present over twenty tests are used to select incumbents in the different clerical occupations. The form of some of these tests have remained unchanged for many years, while the jobs have been remolded in response to the many technological advancements (word processing, automatic data processing, etc.). Secondly, with the signing of the Uniform Guidelines on Employee Selection Procedures (1978) came the necessity to either verify the adequacy of our current examination program by showing that the knowledges, skills and abilities needed for successful job performance were actually those being tested for at the entry level or to develop new selection procedures which met the criteria of the Guidelines.

Our first task was to define clerical occupations and to determine which occupations in the Federal sector met this definition. For the purpose of this study clerical occupations were defined as "those that involve structured work in support of office, business or fiscal operations; performed in accordance with established policies, procedures or techniques; and requiring training, experience, or working knowledges related to the tasks to be performed (U. S. Civil Service Commission, 1976, p. XIX)." With the aid of Standards we were able to come up with thirty-five occupations which met this criteria. The final list of these occupations appear in appendix A. Occupations requiring cryptographic and/or financial skills were not included in this set of occupations for study.

Our next step was to develop a work survey instrument which could be used to gather job information across occupations. We needed a single instrument which incorporated the work behaviors of all thirty-five occupations. Luckily it was not necessary to start from scratch, for there already existed several clerical inventories (Gandy and Maier (1979), state of Minnesota, state of Connecticut, and the Psychological Corporation). These existing inventories, along with Classification Standards for the occupations, previous job analysis studies and the Dictionary of Occupational Titles were consulted in the development of the survey. From these sources we were able to gather a listing of work behaviors or tasks which seemed complete. This list of tasks was reviewed by psychologists at OPM and edited to eliminate duplications. A draft listing of work behaviors by occupation was then given to between three and five job incumbents and/or supervisors in each of the thirty-five occupations for review. They examined the list for completeness, clarity and organization. The suggestions of these reviewers were then incorporated into the final form of the Work Survey for Clerical Occupations.

The final inventory is divided into five parts and contains a listing of 174 task statements or work behaviors, divided among 14 duties. The first section contains questions concerning the incumbents' background. It includes questions on race, sex, ethnicity, education, geographic location and job series, etc. Information gathered here will be used to describe our sample.

The next five sections are concerned with incumbent judgments about their jobs. Here, employees are first asked to review all the tasks and check those they perform on the job. Next, they must rate the tasks they perform on a relative time spent scale - a rating of the amount of time they spent on each task they performed over the past six months compared to all other tasks they perform. The tasks must then be rated on a relative importance scale - the importance of each task they performed in the last six months compared to all other tasks they performed in the same time period. Each incumbent must then complete the above ratings at the duty level.

Tasks were rated on a seven point scale ranging from 1 = very much below average, to 7 = very much above average, as recommended by Morsh and Archer (1967), because of the greater reliability and precision of a seven point scale. A list of equipment used in these occupations was also included in the inventory. Here, incumbents were asked to check off the equipment they use on the job. In this section, as in the task listing, space was alloted for the incumbent to make additions where necessary.

The final part of the inventory enlists the participants' impression about the inventory's clarity, organization and coverage.

At this point the Clerical Work Behavior Inventory has been administered to a sample of job incumbents in grades 2 through 5. This grade span was chosen because entry into these series is usually at the 2 level with an average full performance level of 5.

Initially, it was planned to sample 200 incumbents from each of the thirty-five clerical occupations, but after determining the total population for each occupation this was not considered feasible. The sampling strategy decided upon took into consideration the population of each occupational series and in so doing attempted to have fair representation of incumbents in each of the thirty-five series.

Steps taken to derive the sample were as follows:

(1) The CPDF (Central Personnel Data File) was used to determine the population size of each occupation.

(2) Sample size by occupation was determined by the above information. The most populous occupations were represented by a larger sample size.

(3) Using CPDF data, the largest employing agencies for each of the occupations was determined.

(4) Knowing the size of the sample needed for each occupation, we took the five agencies employing the largest number of incumbents

for each series and weighted the number of incumbents from these agencies to the total size of the sample needed for each occupation.

(5) Number of people was chosen in order to obtain the largest grade range and geographical spread for each occupation and each agency within each occupation.
In occupations with less than 200 incumbents a sample as close to 100 percent as possible was drawn.

It is hoped that the data analysis will show the degree of overlap between these occupations and aid in the development of new selection procedures or support our current examination procedures.

REFERENCE NOTES

Gandy, J. A. and Maier, W. Utah clerical link-up study: Comparison of Federal and State jobs. (PRR 79-1), Washington, D. C.: U. S. Office of Personnel Management, Personnel Research and Development Center, January, 1979.

Morsh, J. E. and Archer, W. B. Procedural guide for conducting occupational surveys in the United States Air Force. (Tech. Rep. PRL-TR-67-11), Lackland Air Force Base, Texas; Personnel Research Laboratory, 1967.

Sapinkopf, R. C. Proposal to develop new selection procedures for clerical occupations. Washington, D. C.: U. S. Office of Personnel Management, Personnel Research and Development Center, 1979.

REFERENCES

Uniform Guidelines on employee selection procedures. Federal Register, 1978, 43, 38290-38315.

U. S. Civil Service Commission. Occupations of Federal white collar workers. Washington, D. C.: Bureau of Personnel Management Information Systems, (Pamphlet 56-12), 1976.

APPENDIX A

CLERICAL OCCUPATIONS IN THE FEDERAL GOVERNMENT

GS Series	Title
134	Intelligence Clerk
203	Personnel Clerk
204	Military Personnel Clerk
301	General Clerk
302	Messenger
304	Information Receptionist
305	Mail and File Clerk
309	Correspondence Clerk
312	Clerk-Steno Reporter
316	Clerk-Dictating Machine Transcriber
318	Secretary
319	Closed Microphone Reporter
322	Clerk-Typist
324	Cold Type Composing Machine Operator
350	Office Machine Operator
351	Printing Clerk
354	Bookkeeping Machine Operator
355	Calculating Machine Operator
356	Data Transcriber
357	Coding Clerk
359	Electric Accounting Machine Operator
382	Telephone Operator

GS Series	Title
385	Teletypist
394	Communication Clerk
998	Claims Clerk
1021	Office Drafter
1046	Clerk-Translator
1087	Editorial Assistant
1531	Statistical Clerk
2091	Sales Store Clerk
2131	Freight Rate Clerk
2132	Travel Clerk
2133	Passenger Rate Clerk
2134	Shipment Clerk
2151	Dispatcher

Dickinson, Richard W., Texas A&M University, College Station, Texas.
(Wed. A.M.)

CODAP 80: The New Occupational Analysis Computer System

For the past 2 1/2 years, a new version of the job analysis computer software system, the Comprehensive Occupational Data Analysis Programs (CODAP), has been under development at the Occupational Research Division, Industrial Engineering Department, Texas A&M University. The time for the system's initial release is almost at hand. The new system represents a radical departure from existing job analysis computer software in that particular attention has been given to making communication between the user and the system as "friendly" as possible, while at the same time maintaining the system's viability through its flexibility in processing an occupational database. The new system's present capabilities will be discussed, with particular emphasis being given to the future potential the system represents in the analysis of occupational information.

CODAP80: THE NEW OCCUPATIONAL ANALYSIS COMPUTER SYSTEM

Richard W. Dickinson
Occupational Research Division
Industrial Engineering Department
Texas A&M University

INTRODUCTION

For the past 2 1/2 years, CODAP80, a new job analysis computer software system, has been under development at the Occupational Research Division of the Industrial Engineering Department at Texas A&M University. The new system represents a radical departure from existing job analysis computer software in that particular attention has been given to making communication between the user and the system as convenient and "friendly" as possible, while at the same time providing the job analyst with a database processing tool flexible enough to meet both present and future applications in the analysis of occupational information.

CODAP80

CODAP80 is a sophisticated software system for processing occupational data collected with job inventories. It was designed with the particular needs of the job analyst in mind. Many aspects of the system will satisfy demands unique to their work, and much of the system's terminology is oriented toward them. In addition to its specialized job analysis characteristics, the data handling and analysis features of the system provide general database management capabilities.

The CODAP80 computer system provides the job analyst with the tools necessary to flexibly investigate occupational information pertaining to classification of employees, training emphasis, work-load distribution, promotion standards, profile analysis, item analysis and personnel assignment. For those questions routinely asked of occupational data users will find that CODAP80 represents a very convenient and powerful mechanism for providing answers.

Information collected with job inventoreis can usually be conceptualized in the form of a 2-Dimensional matrix (see Figure 1), with the columns of the matrix representing incumbents (or workers) and the rows of the matrix representing the individual data items collected from each incumbent. These data items (or variables) may consist of background information (such as incumbent age, race, sex, job title, etc.) or may consist of incumbent responses to time spent on work tasks, equipment usage or knowledge, skill and ability indicies. Since CODAP80 makes very few assumptions about the data that has been collected on incumbents, few, if any, restrictions are imposed as to the type of job related information that may be surveyed.

Processing of the occupational database is performed through the use CODAP80's easy to learn, English-like language. Specific procedures in the language allow the user to conveniently summarize both rows and columns of a 2-Dimensional occupational database, as well as provide report displa particularly suited to occupational analysis and interpretation.

CODAP80 and Hierarchical Clustering

When investigating occupational information collected with job inventories, job analysts frequently find it useful to have the incumbent workers surveye grouped as a function of their similarity on some job related dimension (oft this dimension is work time across tasks, but could just as well be any other job related index, such as equipment usage). In this regard, CODAP80 is of particular utility owing to its capability to produce statistical summaries and report displays based on criteria established through the use of the system's powerful hierarchical clustering procedure.

Classification of incumbent workers can proceed in a quantitative and systematic fashion when grouping or clustering techniques are applied to an occupational database. Interpretation of the results of a cluster operation, though, is often difficult. CODAP80 eases the job analysts' interpretive burden in two ways:

1) As part of the output generated through execution of CODAP80's hierarchical clustering procedure, a pictoral dendogram (or tree-diagram) is printed that visually illustrates the grouping process (see Figure 2). In effect, the job analyst is provided with a "picture" of how the incumbent workers were combined as a function of criterion homogeneity.

2) CODAP80 "remembers" the collapsing sequence as incumbent workers are combined during the clustering process. This "memory" of the grouping process allows very convenient reference of incumbent data. The user need only specify to CODAP80 (through the use of the system's English-like language) the cluster groupings of interest, and the system automatically directs access and summarization of the database as a function of this.

The capability to automatically direct database processing based on results supplied from a cluster operation provides CODAP80 with a distinct advantage over other software packages when analyzing occupational information (particularly when investigating information pertaining to classification).

CODAP80 and Report Generation

CODAP80 provides the job analyst with the ability to easily process an occupational database and then display the processed information in a manner most amenable to interpretation. The vehicle of communication between the job analyst and CODAP80 is an English-like language that conveys to the system the processing and report displays desired.

Suppose, for example, the job analyst, after study of the tree-diagram displayed in Figure 2, desired to investigate the background characteristics of those incumbents defined by clustering to be in groups 73 and 67 (enclosed in brackets, Figure 2). Statistics supplied with the output of the tree-diagram indicate the two cluster groups to be sizable, with both demonstrating relatively high between and within homogeneity. To produce a listing that would conveniently display the background (or history) information of the incumbents in cluster groups 73 and 67, the job analyst need only specify the following CODAP80 language statements:

```
BEGIN DATABASE EXECUTE.
PRINT COLUMNS (G73 G67) NOREMARKS NOSUMMARY AVGP STDP SUM N /
      ROWS (H1-H9)
      HEADING:='LISTING-OUT HISTORY VARIABLES H1-H9 FOR THE INCUMBENTS'
               'IN CLUSTER GROUPS G73 AND G67'
               'OUTPUT IS IN HIERARCHICAL SEQUENCE ORDER'.
END.
```

The output generated from execution of the above statements is displayed in Figures 3 and 4. Immediately apparent from examination of the output is the disparity in job title (history variable 5) representation between groups 73 and 67. Stemming from this point, the job analyst may then use CODAP80 to provide information such as those tasks performed by the incumbents in the two groups and the distribution of their work time across those tasks.

Another topic of interest to the job analyst could be average per task pay rate of the individuals performing the tasks. For example, the job analyst may want to investigate the pay rate characteristics per task of the incumbents performing supervisory functions as opposed to incumbents performing lower-level, delegated functions. To define the tasks of interest, perform the desired calculations and produce the report, the job analyst would specify the following CODAP80 source language statments:

```
BEGIN DATABASE EXECUTE.
SELECT ROWS MODULE1 (T105-T123)
            'TASKS INVOLVED IN PERFORMING SUPERVISOR FUNCTIONS';
       ROWS MODULE2 (T187-T209)
            'TASKS INVOLVED IN OPERATING ON-LINE HARDWARE';
       ROWS MOD1MOD2 (T105-T123, T187-T209)
            'COMBINING MODULE1 & MODULE2 FOR CONVENIENCE'.
AVALUE ROWS MOD1MOD2 FOR (INCUMBENTS) USING H9
       AVGPAYRATE := AVGP 'AVERAGE PAY RATE FOR INCUMBENTS PERFORMING TASKS';
       STDPAYRATE := STDP 'STD OF PAY RATE FOR INCUMBENTS PERFORMING TASK';
       NPAYRATE   := N    'NUMBER OF OBSERVATIONS IN CALCULATING AVERAGE AND
PRINT  ROWS (MODULE1 MODULE2) NOSUMMARY AVGP STDP SUM N
       COLUMNS (AVGPAYRATE1 STDPAYRATE1 NPAYRATE1)
       HEADING    := 'OUTPUT LISTING SUPERVISORY AND NONSUPERVISORY TASKS'
                     'AS WELL AS THE N, STD AND AVERAGE PAY RATE OF THE'
                     'INCUMBENT PERFORMING THE TASKS'.
END.
```

The output generated from execution of the above CODAP80 commands is displayed in Figures 5 and 6. In addition to the average pay rate per task for those incumbents performing the task, the standard deviation and number of observations has also been calculated and displayed. As would be expected, the output clearly shows that the incumbents performing supervisory tasks are generally paid better than the incumbents performing nonsupervisory tasks.

CODAP80's English-like language is easy to learn and use. The language statements described in this paper represent only a small fraction of the leverage CODAP80 can supply to occupational analysis. Provided with CODAP80, the job analyst can actively investigate an occupational database in both a subtle and powerful manner.

Figure 1. Conceptual representation of hypothetical occupational database.

	I1	I2	I3	I4	I5	I6	I7
H1	1	2	1	1	1	2	1
H2	23	*	*	41	27	19	53
H3	2	11	16	19	3	1	30
H4	5	7	3	2	4	1	6
T1	11	0	0	11	24	64	36
T2	11	0	43	44	24	9	64
T3	22	20	57	0	18	9	0
T4	56	50	0	22	0	18	0
T5	0	30	0	22	35	0	0
S1	*	*	*	2	*	*	2
S2	1	*	3	2	1	1	1
S3	2	2	3	*	1	1	*
S4	1	2	*	2	*	2	*
S5	*	1	*	1	3	*	*

STUDY ID

DATABASE

REMARKS

H1 SEX
H2 AGE
H3 YEARS ON JOB
H4 INCUMBENT ID
T1 SUBDUE VIOLENT INMATES
T2 SHAKE DOWN INMATES
T3 SHAKE DOWN VISITORS
T4 ESCORT INMATES
T5 TESTIFY IN COURT
S1 SECONDARY - SUBDUE VIOLENT INMATES
S3 SECONDARY - SHAKE DOWN INMATES
S4 SECONDARY - SHAKE DOWN VISITORS
S4 SECONDARY - ESCORT INMATES
S5 SECONDARY - TESTIFY IN COURT

Figure 2. First page of output displaying a pictoral dendogram (or tree-diagram) illustrating the grouping process during hierarchical clustering.

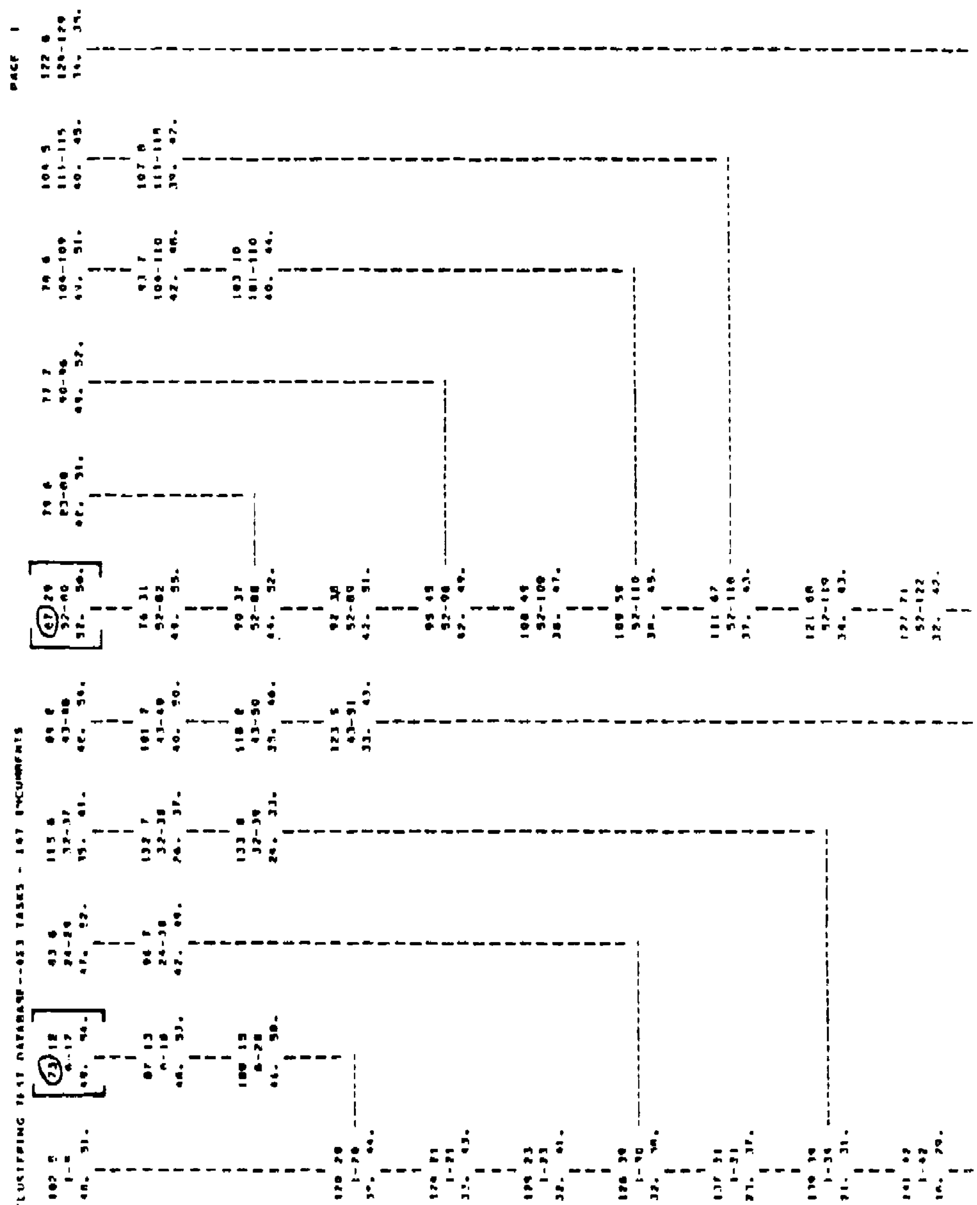


Figure 3. Output generated by CODAP80 displaying the background (history) information of the incumbents in cluster group G73.

PAGE - 1

STUDY ID DATABASE
LISTING-OUT HISTORY VARIABLES HI-H9 FOR THE INCUMBENTS
IN CLUSTER GROUPS G73 AND G87
OUTPUT IS IN HIERARCHICAL SEQUENCE ORDER

H - 1 CASE NUMBER
H - 2 AGE
H - 3 SEX
H - 4 ETHNIC GROUP
H - 5 JOB TITLE
H - 6 DEPARTMENT IN WHICH YOU WORK
H - 7 BUILDING WHERE JOB IS LOCATED
H - 8 HOURS WORK EACH WEEK
H - 9 YEARLY GROSS PAY RATE

G - 73	H - 1	H - 2	H - 3	H - 4	H - 5	H - 6	H - 7	H - 8	H - 9
1 - 6	1089.00	28.00	1.00	1.00	21.00	6.00	8.00	40.00	13900.00
1 - 7	1090.00	29.00	1.00	1.00	20.00	6.00	8.00	40.00	17000.00
1 - 8	1097.00	33.00	1.00	1.00	20.00	6.00	8.00	40.00	14000.00
1 - 9	1096.00	23.00	2.00	1.00	22.00	6.00	8.00	40.00	11800.00
1 - 10	1102.00	27.00	2.00	1.00	22.00	6.00	8.00	40.00	12600.00
1 - 11	1103.00	32.00	1.00	1.00	21.00	6.00	8.00	40.00	17250.00
1 - 12	1055.00	31.00	1.00	1.00	20.00	6.00	8.00	24.00	9200.00
1 - 13	1104.00	43.00	1.00	1.00	20.00	6.00	8.00	40.00	10148.00
1 - 14	1100.00	28.00	2.00	1.00	21.00	6.00	8.00	40.00	13500.00
1 - 15	1095.00	27.00	1.00	3.00	22.00	6.00	8.00	40.00	12000.00
1 - 16	1007.00	26.00	1.00	1.00	21.00	1.00	1.00	40.00	13000.00
1 - 17	1003.00	31.00	1.00	1.00	21.00	1.00	1.00	40.00	13000.00
AVGE	1078.42	29.83	1.25	1.33	20.92	5.17	6.83	38.67	13704.00
STDV	36.67	5.01	.45	1.15	.79	1.95	2.72	4.62	2356.14
SUM	12941.00	358.00	15.00	14.00	251.00	62.00	82.00	464.00	164448.00
N	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00	12.00

Copy available to DTIC does not permit fully legible reproduction

Figure 4. Output generated by CODAP80 displaying the background (history) information of the incumbents in cluster group G67.

PAGE - 2

STUDY ID DATABASE
LISTING-OUT HISTORY VARIABLES H1-H9 FOR THE INCUMBENTS
IN CLUSTER GROUPS G73 AND G67
OUTPUT IS IN HIERARCHICAL SEQUENCE ORDER

H - 1 CASE NUMBER
H - 2 AGE
H - 3 SEX
H - 4 ETHNIC GROUP
H - 5 JOB TITLE
H - 6 DEPARTMENT IN WHICH YOU WORK
H - 7 BUILDING WHERE JOB IS LOCATED
H - 8 HOURS WORK EACH WEEK
H - 9 YEARLY GROSS PAY RATE

G - 67	H - 1	H - 2	H - 3	H - 4	H - 5	H - 6	H - 7	H - 8	H - 9
I - 52	1064.00	30.00	1.00	1.00	15.00	5.00	6.00	30.00	11500.00
I - 53	1063.00	29.00	2.00	1.00	15.00	5.00	6.00	30.00	10000.00
I - 54	1252.00	28.00	1.00	1.00	15.00	11.00	13.00	40.00	13900.00
I - 55	1075.00	32.00	1.00	1.00	18.00	12.00	1.00	40.00	15900.00
I - 56	1255.00	27.00	1.00	1.00	14.00	11.00	13.00	40.00	17600.00
I - 57	1048.00	29.00	1.00	1.00	14.00	4.00	8.00	40.00	17000.00
I - 58	1114.00	20.00	1.00	1.00	13.00	6.00	9.00	20.00	10950.00
I - 59	1109.00	29.00	1.00	1.00	17.00	6.00	14.00	40.00	9336.00
I - 60	1084.00	25.00	1.00	1.00	16.00	6.00	8.00	40.00	11016.00
I - 61	1037.00	34.00	1.00	1.00	6.00	1.00	1.00	40.00	22900.00
I - 62	1029.00	30.00	2.00	1.00	6.00	1.00	1.00	40.00	29000.00
I - 63	1042.00	28.00	2.00	1.00	10.00	1.00	1.00	40.00	20000.00
I - 64	1041.00	30.00	1.00	1.00	10.00	1.00	1.00	40.00	20100.00
I - 65	1016.00	32.00	1.00	1.00	5.00	1.00	1.00	40.00	23800.00
I - 66	1044.00	35.00	1.00	1.00	7.00	1.00	1.00	40.00	19000.00
I - 67	1025.00	32.00	1.00	1.00	5.00	1.00	1.00	40.00	21900.00
I - 68	1031.00	33.00	2.00	1.00	8.00	1.00	1.00	40.00	24060.00
I - 69	1046.00	49.00	1.00	1.00	7.00	1.00	1.00	40.00	14020.00
I - 70	1045.00	29.00	1.00	1.00	9.00	1.00	1.00	40.00	22200.00
I - 71	1093.00	27.00	1.00	1.00	10.00	1.00	1.00	40.00	19000.00
I - 72	1019.00	34.00	1.00	1.00	6.00	1.00	3.00	40.00	22290.00
I - 73	1027.00	33.00	1.00	1.00	10.00	1.00	1.00	40.00	18400.00
I - 74	1077.00	37.00	2.00	1.00	15.00	10.00	12.00	40.00	14400.00
I - 75	1400.00	39.00	2.00	1.00	6.00	10.00	12.00	40.00	18200.00
I - 76	1069.00	31.00	1.00	1.00	11.00	7.00	10.00	40.00	17600.00
I - 77	1028.00	33.00	1.00	1.00	5.00	2.00	2.00	40.00	23000.00
I - 78	1076.00	27.00	2.00	1.00	10.00	12.00	1.00	40.00	20000.00
I - 79	1025.00	46.00	1.00	1.00	6.00	1.00	1.00	40.00	22000.00
I - 80	1078.00	28.00	2.00	1.00	14.00	12.00	1.00	40.00	17000.00
AVGE	1077.18	32.10	1.28	1.00	10.34	4.66	4.55	38.62	18052.53
STDD	85.03	5.41	.45	.00	3.98	4.25	4.75	4.41	4261.00
SUM	31744.00	931.00	37.00	29.00	300.00	135.00	132.00	1120.00	523524.00
N	29.00	29.00	29.00	29.00	29.00	29.00	29.00	29.00	29.00

Figure 5. Output generated by CODAP80 displaying supervisory tasks as well as pay rate statistics of the incumbents performing these tasks.

STUDY ID CATABASE
OUTPUT LISTING SUPERVISORY AND NONSUPERVISORY TASKS
AS WELL AS THE N, STD AND AVERAGE PAY RATE OF THE
INCUMBENTS PERFORMING THE TASKS

AVGPAYRATE1 AVERAGE PAY RATE FOR INCUMBENTS PERFORMING TASK
STDPAYRATE1 STD OF PAY RATE FOR INCUMBENTS PERFORMING TASK
NPAYRATE1 NUMBER OF OBSERVATIONS IN CALCULATING AVERAGE AND STD

NUMBER T	TASKS INVOLVED IN PERFORMING SUPERVISORY FUNCTIONS	AVGPAYRATE1	STDPAYRATE1	NPAYRATE1
T - 105	SUPERVISE PERSONNEL DEVELOPING AND DELIVERING IN-HOUSE TRAINING AND EDUCATIONAL ACTIVITIES	17036.15	5833.11	13.00
T - 106	SUPERVISE PERSONNEL DOING SYSTEMS DESIGN ACTIVITIES	21157.23	3819.51	21.00
T - 107	SUPERVISE PERSONNEL DOING SYSTEMS DEVELOPMENT ACTIVITIES	20916.75	3667.53	24.00
T - 108	SUPERVISE PERSONNEL DOING SYSTEMS MAINTENANCE ACTIVITIES	20274.25	3846.02	24.00
T - 109	SUPERVISE PERSONNEL OPERATING ON-LINE HARDWARE	15512.94	4170.31	34.00
T - 110	SUPERVISE PERSONNEL OPERATING OFF-LINE HARDWARE	15031.82	3705.83	22.00
T - 111	SUPERVISE PERSONNEL OPERATING MAIN CONSOLE	14670.78	4141.56	23.00
T - 112	SUPERVISE PERSONNEL WORKING WITH STORAGE MEDIA	14496.66	4456.88	24.00
T - 113	SUPERVISE PERSONNEL CREATING COMPUTER PROGRAMS	20337.33	3871.33	33.00
T - 114	COUNSEL EMPLOYEES WITH PERSONAL PROBLEMS RELATED TO WORK	17291.26	4870.14	38.00
T - 115	MAKE THE FINAL DETERMINATION CONCERNING WHICH HARDWARE TO PURCHASE	19398.00	4803.10	10.00
T - 116	DETERMINE WHETHER OR NOT EMPLOYEES ARE BEING USED IN THE MOST EFFICIENT MANNER	18587.09	4542.41	31.00
T - 117	CONTACT OTHER INSTALLATIONS TO SEE ABOUT RENTING OR PURCHASING SOME OF THEIR HARDWARE OR SOFTWARE	18243.25	4635.77	24.00
T - 118	ANALYZE THE UTILIZATION OF COMPUTER EQUIPMENT TO DETERMINE IF IT IS BEING USED MOST EFFICIENTLY	18545.06	4506.95	30.00
T - 119	APPROVE/DISAPPROVE RECOMMENDATIONS CONCERNING CHANGES IN THE DIVISION	19599.33	4720.52	12.00
T - 120	SUPERVISE PERSONNEL PERFORMING SECRETARIAL FUNCTIONS	17860.00	4513.97	10.00
T - 121	SUPERVISE MAINTENANCE WORK OF CUSTOMER ENGINEER	15397.20	4490.98	10.00
T - 122	SUPERVISE JOB PRODUCTION SCHEDULES	16728.00	4311.05	20.00
T - 123	SUPERVISE TESTS AND EVALUATIONS OF COMPUTER PROGRAMS	18985.79	4158.91	20.00
AVGP		17896.29	4369.88	22.69
STDP		2151.89	515.35	8.59
SUM		340029.44	83027.69	431.00
		19.00	19.00	19.00

Figure 6. Output generated by CODAP80 displaying nonsupervisory tasks as well as pay rate statistics of the incumbents performing these tasks.

STUDY ID DATABASE
OUTPUT LISTING SUPERVISORY AND NONSUPERVISORY TASKS
AS WELL AS THE N, STD AND AVERAGE PAY RATE OF THE
INCUMBENTS PERFORMING THE TASKS

AVGPAYRATE1 AVERAGE PAY RATE FOR INCUMBENTS PERFORMING TASK
STDPAYRATE1 STD OF PAY RATE FOR INCUMBENTS PERFORMING TASK
NPAYRATE1 NUMBER OF OBSERVATIONS IN CALCULATING AVERAGE AND STD

MODULE2	TASKS INVOLVED IN OPERATING ON-LINE HARDWARE	AVGPAYRATE1	STDPAYRATE1	NPAYRATE1
T - 187	RECEIVE REPORTS OF EQUIPMENT PROBLEMS	15600.16	4944.89	49.00
T - 188	REPORT EQUIPMENT PROBLEMS	14014.84	4877.98	74.00
T - 189	COMPLETE SPECIAL REQUEST FORMS TO OBTAIN COMPUTING SERVICE	14551.29	4154.20	28.00
T - 190	OPERATE CDC 8231 CONSOLE	12983.33	4327.05	24.00
T - 191	OPERATE CDC 1700 CONSOLE	12547.11	3682.11	18.00
T - 192	OPERATE DEC 10 CONSOLE	13342.50	3942.34	20.00
T - 193	OPERATE PAPER TAPE READER	13177.52	4401.55	25.00
T - 194	OPERATE PAPER TAPE PUNCH	14282.78	4290.85	18.00
T - 195	OPERATE DISC DRIVE	14463.02	5083.84	43.00
T - 196	OPERATE TAPE DRIVE	13488.83	4744.55	60.00
T - 197	OPERATE CARD PUNCH	13370.72	4670.66	47.00
T - 198	CLEAR CARD PUNCH JAMS	12864.98	4571.16	45.00
T - 199	OPERATE VERSATEC PRINTER	13280.00	4023.86	21.00
T - 200	CLEAR JAMS AND MISREADS ON CARD PRINTER	12670.93	4053.21	30.00
T - 201	OPERATE CARD READER	13658.19	4879.75	74.00
T - 202	OPERATE PLOTTER	13551.11	4637.95	36.00
T - 203	INFORM THE USER (OR OPERATOR) OF NECESSARY STEPS TO CORRECT PROBLEMS AT HJE SITE	13786.59	4241.71	87.00
T - 204	REVIEW SPECIAL REQUEST FORMS TO DETERMINE IF COMPLETE	13630.86	3817.07	29.00
T - 205	LOCATE APPROPRIATE PERSONNEL TO SOLVE SPECIFIC PROBLEMS	14441.97	4851.40	80.00
T - 206	RUN THE DEADSTART DATA DECK AT REMOTE JOB ENTRY SITE	13022.42	4671.21	33.00
T - 207	OPERATE GRAPHIC DISPLAY UNIT	14822.73	5676.73	22.00
T - 208	CONNECT TERMINALS TO NEW OR TEMPORARY SITE IN COMMUNICATIONS AREA	16642.50	6923.91	16.00
T - 209	INSTALL TERMINALS AT REMOTE LOCATIONS	15865.00	4635.96	12.00
AVGP		13914.38	4526.03	36.96
STDP		1050.10	467.04	19.84
SUM		320030.88	104098.69	850.00
N		23.00	23.00	23.00

Dodd, B. T., Royal Navy School of Educational and Training Technology, HMS Nelson, Portsmouth, England. (Fri. A.M.)

Current Directions in Selection Research for the Royal Navy

The computer assisted personnel selection system reported at MTA 1980 is moving into the procurement stage although the first implementation will not include most of the self-correcting mechanisms built into the prototype. Doubts concerning the volume of recruiting business required in the future have curtailed the money available.

Counselling, the representation of military technical life, and aides to self selection have been of major concern in view of the small intakes required and the consequent disturbances when selectees decline to join or leave prematurely.

Microprocessor-based tests are being considered in some areas but it is not expected that paper testing will be supplanted.

Short courses of a few days' length at the Royal Marine Commands Training Center are being evaluated as a selection device in an attempt to reduce the leaving rate due to unexpected physical demands and life styles.

CURRENT DIRECTIONS IN SELECTION FOR THE ROYAL NAVY

Selection of Royal Navy Ratings, Women's Royal Naval Service, and Royal Marine Other Ranks is carried out by permanent 'Careers Advisors' who have retired from the Service. At this stage they are trained by the Senior Psychologist (Naval), SP(N), who thereafter supplies them with tests and other selection instruments and sees they are properly used. Wastage, advancement and return of service are monitored by SP(N) who can also track the success of individual recruiters.

As reported at MTA in 1980, SP(N) has developed a computer model of the selection system that would be possible if each recruiter could link to a central database to learn how attractive a particular applicant would be in each of the jobs for which he was eligible on each of the entry dates for which there was a vacancy.

This somewhat idealised system has been pruned by economic circumstances to leave a rather different proposal. What is going to be installed is a recruiting management system with terminals only at regional headquarters. There is no money for the recruiter to be equipped like an airline booking clerk.

What will be new about the recruiting management system will be an estimate of the Recruiting Test score that an applicant needs in order to have a good chance of being allocated a place. The RN has a few jobs difficult to fill with suitable applicants, but for most branches, eligible applicants exceed requirements. If one's Recruiting Test scores is below the estimated cutting score, the chances of a place are almost nil. This is expected to save on recruiting resources.

Yet a further measure being taken to stem the flood of marginal applicants is to make sure that enquirers have a chance to realise what they are contemplating. At present, job preview provisions range from a recruiting leaflet to a video cassette. For Royal Marines there is a potential recruit course lasting some two days at which the curious may sample training in the Royals and the training staff may assess just who should be offered the limited places.

In between the advertising leaflet and the two-day potential recruits course there are explanatory booklets, charts and posters. These are being overhauled and assembled into folios of texts and photographs to portray what it will be like to live and work in a naval training environment, one folio for each trade. Whilst the surfeit lasts, this emphasis on job information will be the principal method of deterring those who might later discover that they had made a wrong choice.

For the near future, at least, selection testing will continue to be conducted using pencils and booklets, answers being marked by hand against a key. A structured biographical interview lasting about three-quarters of an hour is the main source of non-test information although the interviewer's judgments are later recorded on a seven-dimensional scaling form. Experiments with a job disposition questionnaire and a job knowledge index are in progress although it will be some years before follow-up data will be available to validate those devices.

As a policy for near future action, the naval personnel research effort will devote rather more resources to seeking proper criterion data on which to validate selection practices. One line of work is attempting to capture significant job knowledge in ways which allow computerised end-of-training testing. Such testing is expected to yield useful data on the absorption and retention of job knowledge under the current training regime. The art of criterion test development and its comprehensive practice are not conspicuous in naval training.

With the pressure to conduct more training at sea, job performance or job knowledge testing by shipborne systems is becoming a promising line of development which will benefit both training and selection authorities.

AD P001308

AN EVALUATION OF THE FAIRNESS OF THE FLIGHT APTITUDE SELECTION TEST (FAST)

John A. Dohme, Ph.D.
Research Psychologist
US Army Research Institute
Fort Rucker Field Unit

The concept "test fairness" has developed only recently. A major impetus in the development and application of the concept has come from the publication of the Uniform Guidelines on Employee Selection Procedures (UGES) in 1978. The UGES are interpreted as mandating the use of a regression model in evaluating test fairness. A technique was developed utilizing a regression model to evaluate the fairness of the Flight Aptitude Selection Test (FAST) for the groups identified by the UGES: Blacks, American Indians, Asians, Hispanics, Caucasians and females. The regression of FAST scores on overall grades in the Initial Entry Rotary Wing (IERW) course was performed for each of the above groups in comparison with the majority group. Available population sizes were considered too small to permit a conclusive fairness evaluation at this time. The fairness evaluation will be repeated semiannually until minority population sizes permit sufficient power to perform a definitive analysis.

AN EVALUATION OF THE FAIRNESS OF THE FLIGHT APTITUDE SELECTION TEST (FAST)

"Fairness" as a criterion for the evaluation of a test or other selection procedure is a relatively new concept. The concept has evolved from the technology of test validation to answer the question, "Is this test/procedure valid for the selection of minority as well as majority applicants?" Appropriate methodology for the evaluation of fairness is currently a matter for debate in the technical literature (Ledvinka, 1979). A major impetus for the development of fairness methodologies was the publication of Guidelines on Employee Selection Procedures in 1970 by the Equal Employment Opportunity Commission. In fact, the most current version, the Uniform Guidelines on Employee Selection (UGES) (1978), noted that, "The concept of fairness or unfairness of selection procedures is a developing concept, (14B(8))." Since this technology is still developmental, this paper will review the rationale and precedence for the FAST fairness evaluation in some detail.

Technical standards for performing a fairness evaluation are addressed by both professional and government agencies. The American Psychological Association (APA) publication, Principles for the Validation and Use of Personnel Selection Procedures (1975), discusses both technical and ethical implications of the choice of methodology in fairness research designs. The government publication referenced above, Uniform Guidelines on Employee Selection Procedures (UGES) published in 1978, which is a codified position agreed upon by the US Civil Service Commission, the Department of Justice, the EEOC, and the Department of Labor falls under the scope of Title VII of the 1964 Civil Rights Act and, for that reason, carries the impact of law.[1] Furthermore, the current version of the UGES was reviewed by the APA prior to publication, thus, it is a synthesis of professional and governmental guidance in the technical and ethnical and legal aspects of fairness research designs. For these reasons, this paper will make frequent reference to the UGES.

The UGES define fairness by stating its obverse: "When members of one race, sex, or ethnic group characteristically obtain lower scores on a selection procedure than members of another group, and the differences in scores are not reflected in differences in a measure of job performance, use of the selection procedure may unfairly deny opportunities to members of the group that obtains the lower scores (Section 14B8a)." This definition has clear implications in the design of a fairness research study in that it specifies that fairness should be defined in terms of the bivariate distribution of test (or other selection procedure) scores and job performance scores. Specifically, fairness is demonstrated by coincident regression of job performance scores on test scores for a minority group and the majority group. Fairness does not require that minority performance on the test, or on the job be equal to majority performance but only that the test (or selection procedure) does not over or under predict minority performance vis a vis majority performance.

The UGES do not require routine demonstration of the fairness of a selection procedure for every minority group identified in section 4B (Blacks, American Indians, Asians, Hispanic and Caucasians). Section 14B(8)(b) states: "Where a selection procedure results in an adverse impact on a race, sex, or ethnic group identified in accordance with the classifications set forth in section 4 above and that group is a significant factor in the relevant labor market, the user generally should investigate the possible existence of unfairness for that group if it is technically feasible to do so." In other words, a demonstration of fairness is required only where:

(1) there is evidence of adverse impact as defined in section 4D of the UGES;

(2) that adverse impact affects a group identified in section 4B of the UGES;

(3) the group(s) affected comprise a significant factor in the relevant labor market which is defined in section 15A(1)(c) as constituting more than 2% of the labor force in a "relevant labor area";

(4) it is "technically feasible" to investigate the fairness issue. Technical feasibility is defined in section 14B(8)(c) to include:

(a) sufficient sample sizes to achieve statistical significance;

(b) direct comparability of the samples in terms of the actual jobs performed.

---

[1] At this writing, military personnel in DOD agencies do not fall under the purview of Title VII, thus, may not be legally bound to the UGES. However, the author takes the position that the UGES represent current professional thinking in this technical area, therefore, they provide appropriate guidance independent of their status as law.

The issues raised in paragraphs 1-3 above are empirical questions. They are best answered by descriptive data pertaining to the population of applicants to US Army flight training. The Fort Rucker Field Unit of ARI began an investigation of the selection rates of applicants of the groups identified in section 4B of the UGES. Data were requested from MILPERCEN and RCPAC and a quality check was performed on the data obtained from the master files. Master file data were cross referenced with data in the student pilot's flight folders at the Directorate of Training at Fort Rucker. Taking the black group as an example, master file data were missing for over 78% of the trainees, i.e., 78% of individuals who had entered the flight training course did not appear in the master file. Therefore, it must be concluded that the selection rates prior to 1980 are indeterminate and adverse impact cannot be assessed.

With the advent of the revised FAST test (RFAST) which replaced the earlier form in the field in early 1980, the data collection problem referenced above has been alleviated. The RFAST answer sheet requests information on the sex and ethnic status of applicants. All RFAST answer sheets are sent to ARI, Fort Rucker for machine scoring and storage in the RFAST archives, thus, all the information needed to determine whether or not adverse impact exists will be available at ARI Fort Rucker. Given that it commonly takes more than one year between taking the RFAST and graduation from the 34 week training program, it will be some time before adverse impact can be determined for the RFAST.

In the interim, the conservative assumptions will be made that adverse impact does exist for all the groups identified by Section 4B of the UGES, and that each of those groups constitutes more than 2% of the applicant population. Pursuant to Section 14B(8) of the UGES, a fairness evaluation will be undertaken for each group where it is "technically feasible" to do so. However, the issue of technical feasibility is, like the issue of fairness, a matter of some debate in the technical literature. As noted above, the UGES discuss the issue of technical feasibility with reference to sample size and comparability. In an empirical study of the statistical power associated with various sample sizes, Schmidt, Hunter and Urry (1976) concluded:

> *"This study demonstrates that sample sizes required to produce adequate power in empirical validation studies are substantially larger than has typically been assumed. This finding leads to the conclusion that, from the viewpoint of sample-size requirements, criterion-related validity studies are "technically feasible" much less frequently than is commonly assumed (p. 473)."*

Using the methodology developed by Schmidt, Hunter and Urry (1976) to estimate the sample size required in the present evaluation, and making the liberal assumptions that (1) the true validity of the FAST test is .50, (2) the reliability of the Initial Entry Rotary Wing (IERW) overall grade is .60 and (3) 70% of the applicants to the IERW program are accepted, 128 subjects per group would be required to reach a power of .90 (i.e., to have a 90% probability of rejecting the null hypothesis if it is indeed false). Thus, from the standpoint of the Schmidt, Hunter and Urry (1976) article, it is not technically feasible to perform a fairness evaluation of the FAST until a larger sample of IERW graduates is available.

An earlier section of this research report noted that a revised version of the FAST (the RFAST) is presently being implemented in the field. The version of the FAST being evaluated for fairness in this report has two different forms developed for implementation with commissioned officers and enlisted personnel respectively. Since the two forms differ substantially in content and number of items, the current fairness evaluation must be conducted separately for these two populations. There is only one form of the RFAST which has been developed for use with both populations. Therefore, future fairness evaluations will not require separate commissioned and enlisted samples which will considerably ameliorate the problem of collecting samples large enough to permit a conclusive fairness evaluation.

One key issue in the design of a fairness evaluation study is the choice of a statistical model to guide the minority/majority comparisons. Section 14B8 of the UGES raises the point that the concept fairness is still evolving in the literature. Specifically, the choice of a statistical model has been debated for nearly a decade since the publication of the 1970 version of the EEOC Guidelines (see Cole, 1972; Hunter and Schmidt, 1974; Hunter, Schmidt and Rauschenberger, 1977 and Ledvinka, 1979). The current literature focuses on four models which lead to different operational definitions of fairness/unfairness:

1. The regression model (Cleary, 1968) which states that a test is fair if the regression lines predicting job performance are the same (plus or minus sampling variation) for minority and majority groups.

2. The conditional probability model (Darlington, 1971; Cole, 1973) which states that a test is fair if the probability of being selected is the same for minority and majority group members who are actually capable of satisfactory job performance.

3. The constant ratio model (Thorndike, 1971) which states that a test is fair if its selection ratio for minority and majority groups is the same as the selection ratio using a perfectly valid test (or using the criterion measure itself for selection).

4. The quota model which states that a test is fair if its selection ratio is the same for all minority and majority groups regardless of group performance on the job.

While various authors continue to argue the technical and ethnical merits of these models, it has been pointed out by Ledvinka (1979, p. 552) and by Hunter, Schmidt and Rauschenberger (1977, p. 256) that the UGES clearly specify the regression model as being legally appropriate in the conduct of fairness research. Two UGES passages can be cited to document this point.

> *"When members of one race, sex or ethnic group characteristically obtain lower scores on a selection procedure than members of another group, and the differences in scores are not reflected in differences in a measure of job performance . . . ." (Section 14B8a)."*

"If unfairness is demonstrated through a showing that members of a particular group perform better or poorer on the job, then their scores on the selection procedure would indicate through comparison with how members of other groups perform, the user may either revise or replace the selection instrument in accordance with these guidelines, or may continue to use the selection instrument operationally with appropriate revisions in its use to assure compatibility between the probability of successful job performance and the probability of being selected." (*Section 24B8d*).

There is an additional, independent reason to use the regression model in this fairness evaluation. Of the four models, it alone does not require a "pass through" methodology in which IERW applicants are selected for flight training regardless of their FAST scores. While a pass through methodology is technically appropriate in fairness research, it incurs a substantial increase in attrition rate over the use of an efficacious selection procedure. Given that the training costs in the IERW program exceed $125,000 per trainee, the two costs of a pass through program, higher attrition costs and a reduced output of trainees, could conceivably cost the government millions of dollars per year and lead to an even greater shortfall in aviators in the field.

METHOD

The subjects that comprise the minority/female samples include all IERW program trainees who identified themselves as belonging to one of the groups previously identified in the UGES (Black, Hispanic, Asian, American Indian, female) and for whom both FAST and IERW overall grade (OAG) data were available in US Army Aviation Center (USAAVNC) records. The data collected cover the time span July 1975 to July 1979.

In order to develop the regression comparison procedure and to estimate the fairness of the FAST as a predictor of performance in the IERW Program, a sample of the FAST and OAG scores for majority trainees was selected. During the same time period that scores were monitored for the minority samples described in this report, a random sample of 10% of majority officers and 10% of majority WOCs was drawn from the majority population.

The sample sizes for minority/female and majority commissioned officers and WOCs are presented in Table 1.

The Introduction Section of this paper developed the concept that the evaluation of test fairness requires the comparison of minority/female and majority regression lines. A statistical technique was specifically formulated for this purpose by Gulliksen and Wilks (1950). Additionally, there is precedence for the application of this procedure under the mandate of the UGES (Reilly, Zedeck, and Tenopyr, 1979). The Gulliksen Wilks technique, which was derived from Neyman-Pearson likelihood ratio test theory, tests three null hypotheses sequentially (1950, p. 96):

TABLE 1

DESCRIPTIVE STATISTICS FOR THE MINORITY AND MAJORITY COMMISSIONED OFFICER AND WARRANT OFFICER CANDIDATE (WOC) SAMPLES

GROUP	CATEGORY	SAMPLE SIZE	FAST		OVERALL GRADE (OAG)		CORRELATION FAST WITH OAG	SIGNIFICANCE OF THE CORRELATION
			MEAN	STANDARD DEVIATION	MEAN	STANDARD DEVIATION		
BLACK	OFFICER	22	236.36	53.52	85.14	2.64	.537	$p<.01$
	WOC	14	331.57	26.11	85.20	2.37	.434	NS
HISPANIC	OFFICER	14	263.57	59.44	85.90	2.40	-.204	NS
	WOC	19	329.16	22.81	84.88	2.13	-.123	NS
ASIAN	OFFICER	3	258.00	46.87	86.10	0.79	.668	NS
	WOC	12	346.00	31.29	86.74	1.64	.079	NS
AMERICAN INDIAN	OFFICER	9	310.67	57.50	85.62	3.29	.337	NS
	WOC	14	336.93	28.44	85.68	1.97	.197	NS
FEMALE	OFFICER	11	257.18	61.03	84.83	4.04	.703	$p<.01$
	WOC	18	309.00	8.78	85.74	2.41	.244	NS
MAJORITY	OFFICER	94	293.49	65.14	87.07	2.84	.302	$p<.002$
	WOC	117	347.90	29.17	85.83	2.68	.236	$p<.01$

1. H1 is the hypothesis that the populations from which the samples were drawn have equal standard errors of the estimate (around the least squares regression line).

2. H2 is the hypothesis that the slopes of the population regression lines are the same.

3. H3 is the hypothesis that the Y-intercepts of the regression lines are equal.

In applying the technique, the three hypotheses are tested sequentially starting with H1. If any hypothesis is rejected, hypothesis testing stops and it is concluded that the samples were drawn from different bivariate populations. If all three null hypotheses are retained, then the samples have the same bivariate dispersion, slope and intercept and thus, coincident regression lines.

In applying the Gulliksen Wilks technique to the current fairness evaluation, a significant problem arises because of the small sample sizes currently available for ethnic and female IERW trainees. Gulliksen and Wilks state that their primary purpose is, ". . . to present large-sample tests for the hypotheses considered from the point of view of Neyman-Pearson likelihood ratio test theory (1950, p. 94)." The smallest sample in the Reilly, et. al. (1979) experiments included 45 subjects. A conservative statistician would prefer to have 100 data points in a "large sample" bivariate distribution. However, it is clear that the sample sizes in the current research, which range from a high of 22 Black Officers to a low of 3 Oriental Officers, do not meet the sample size requirement for the Gulliksen Wilks procedure.

A search of the statistics literature produced a regression line comparison procedure which was derived from the analysis of covariance rather than from Neyman-Pearson likelihood ratio theory. Snedecor and Cochran (1967, pp. 432-436) present a procedure which tests the same three sequential hypotheses discussed by Gulliksen and Wilks (1950). This procedure, while it is sensitive to the usual assumptions made by parametric statistics, is not based on the assumption of large sample sizes.

RESULTS

Table 1 presents sample sizes, means, and standard deviations for the Commissioned Officer and WOC samples. In addition, the correlation of the FAST and overall grade for each group and the significance of that correlation coefficient is shown. At least in part because of the small sample sizes of the minority and female samples, only 2 of the 10 correlations attained significance. In both of the majority samples, the FAST proved to be a significant predictor of overall grade despite the restriction in range caused by the prior use of FAST scores as a selection criterion (Commissioned Officers must score at least 155 and enlisted or civilian entry must score at least 300[2] to gain admission to the IERW training program). In reality, the

[2]Since these data were collected, the FAST cutoff score for WOCs was reduced to 270.

restriction of range problem applies only to the WOC samples since very few of the Commissioned Officer applicants score below 155. The lesser restriction of range in the officer sample is the most probable explanation for the generally higher correlations in that group, as contrasted to the WOC samples.

The three hypotheses tested in the fairness evaluation concern the equality of the standard errors of the estimate, the slopes, and the Y-intercepts for the minority/female and majority regression lines. The logic of the hypothesis test procedure requires that the three hypotheses be tested sequentially. That is, the hypothesis of equal dispersion about the common regression line is tested first. If that F-ratio reaches significance, the hypothesis test procedure stops and it is concluded that the two samples are not taken from the same bivariate population. If the F-test for equality of variance about the common regression line is nonsignificant, then the second hypothesis is tested, i.e., the two slopes are compared. Again, if the F-ratio reaches significance, it is concluded that the two regression lines are not the same. If the F-ratio is nonsignificant, then the third hypothesis is tested, i.e., the Y-intercepts (or elevations) of the two regression lines are compared. Again, if the F-ratio reaches significance, it is concluded that the two samples did not come from the same bivariate population. Only if all three hypothesis tests yield nonsignificant F-ratios can it be concluded that the two regression lines are coincident.

Given the very small population sizes available at the time this research was undertaken, it might be misleading to present hypothesis test results. The statistical power, even in the largest minority/majority comparison, is not sufficiently large to ensure rejection of the null hypotheses if they are indeed false. Thus, these data will be retained and the fairness analysis will be repeated biannually until such time as sufficient data are available to perform a conclusive study.

DISCUSSION

As noted previously, the data base for minority and female IERW trainees is not of sufficient size to permit drawing conclusions regarding the fairness of the FAST as a selection device. The purpose of this paper is to develop the rationale and methodology for such a fairness evaluation. Thus, the current discussion will focus primarily on methodological issues.

In accordance with the UGES the fairness of a selection procedure should be determined by reference to the regression of that selection test (or procedure) on job referenced criteria. Section 14B(3) of the UGES notes that training performance is an acceptable criterion under certain conditions:

> *"Where performance in training is used as a criterion, success in training should be properly measured and the relevance of the training should be shown either through a comparison of the content of the training program with the critical or important work behavior(s) of the job(s), or through a demonstration of the relationship between measures of performance in training and measures of job performance. Measures of relative success in training include but are not limited to instructor evaluations, performance samples, or tests."*

The IERW training program clearly meets the conditions specified in 14B(3) by virtue of the content of the training program and the measures of relative success employed as grading procedures. The curriculum of the IERW Program of Instruction (POI) has been developed specifically to train aviators to perform Army aviation missions in the field. Thus, the content of the training program corresponds very closely to the critical work behaviors performed on the job. Training grades are composed of the three components identified in the UGES: Instructor evaluations (Instructor Pilot put-up scores), performance samples (checkrides), and tests (academic examinations). The IERW overall grade which is used as a criterion in this research is a composite of all three evaluation components. In summary, the design of the current fairness evaluation is in accordance with the directives of the UGES.

While the sample sizes for the minority/female groups presented in Table 1 are too small to justify the drawing of inferences to the entire populations of female and minority aspirant aviators, several points warrant discussion. For both Hispanic samples (Officer and WOC), the FAST has a nonsignificant negative correlation with overall grade. Inspection of the scatter diagrams in both cases reveals that, while the general linear trend is positive for the entire sample, two or three outliers with extreme scores unduly influenced the regression line. For example, in the Commissioned Officer sample, the individual with the highest IERW overall grade, 89.35, has an unusually low FAST score, 197. Expressed as standard scores, this individual's overall grade is $z = 1.44$ whereas his FAST is $z = -1.12$. Conversely, the individual with the lowest overall grade, 79.39, has a moderately high FAST score, 313. Expressed as standard scores, overall grade $z = -2.71$ and FAST $z = .83$. If these two individuals are removed from the distribution, the correlation for the remaining 12 individuals is .193. The sensitivity of this correlation coefficient to only two data points demonstrates the inappropriateness of generalizing from the small minority and female samples in the current study.

The purpose of this research effort is to establish an appropriate methodology to evaluate the FAST for fairness. The methodology reviewed in this paper has been programmed for automated computation on a computer. Additionally, a mechanism has been established to collect data on minority/female and majority IERW trainees. As more minority/female trainees complete pilot training, the fairness evaluation will be iteratively performed until sample sizes permit sufficient statistical power to draw conclusions about the fairness of the FAST.

REFERENCES

Cleary, T. A. *Test bias: Prediction of grades of negro and white students in integrated colleges*. Journal of Educational Measurement, 1968, 5, 115-124.

Cole, N. S. *Bias in selection*, 1972, American College Testing Program, Iowa City, IA.

Cole, N. S. *Bias in selection*. Journal of Educational Measurement, 1973, 10, 237-255.

Cronbach, L. J. *Essentials of psychological testing*. New York: Harper and Brothers, 1960.

Darlington, R. B. *Another look at "culture fairness."* Journal of Educational Measurement, 1971, 8, 71-82.

Guidelines on employee selection procedures. Federal Register, 1970, 35, 149, 12333-12336.

Gulliksen, H. *Theory of mental tests*. New York: John Wiley and Sons, 1950.

Gulliksen, H., and Wilks, S. S. *Regression tests for several samples*. Psychometrika, 1950, 15, 91-114.

Hunter, J. E., Schmidt, F. L., and Rauschenberger, J. M. *Fairness of psychological tests: Implications of four definitions for selection utility and minority hiring*. Journal of Applied Psychology, 1977, 62, 245-260.

Ledvinka, James. *The statistical definition of fairness in the federal selection guidelines and its implications for minority employment*. Personnel Psychology, 1979, 32, 551-562.

McMullen, R. L. *An evaluation of selection rates for minority and female applicants to the US Army initial entry rotary wing flight training program*. US Army Research Institute, Ft Rucker Field Unit, Working Paper, 1981.

Principles for the validation and use of personnel selection procedures, 1975, Division 14 (Division of Industrial and Organizational Psychology), American Psychological Association.

Reilly, R. R., Zedeck, S., and Tenopyr, M. L. *Validity and fairness of physical ability tests for predicting performance in craft jobs*. Journal of Applied Psychology, 1979, 64, 262-274.

Schmidt, F. L., Hunter, J. E., and Urry, V. W. *Statistical power in criterion-related validity studies*. Journal of Applied Psychology, 1976, 61, 473-485.

Snedecor, G., and Cochran, W. G. *Statistical methods* - (6th ed.). Ames: Iowa State University Press, 1967.

Thorndike, R. L. *Concepts of culture fairness*. Journal of Educational Measurement, 1971, 8, 63-70.

Uniform guidelines on employee selection procedures. Federal Register, 1978, 43, 166, August 25, 1978.

Doorley, Richard D., US Army Military Personnel Center, Alexandria, Virginia. (Wed. A.M.)

The "New" Army Officer Evaluation Reporting System

The latest Officer Evaluation Reporting System (OERS) represents the most substantive change in officer evaluation concept and philosophy since World War II. It incorporates several new features which have not been included in previous officer evaluation reporting systems: rated officer participation, senior rater concept and a senior rater profile. Participation by the rated officer in the evaluation process provides additional information from the rated officer's point of view to rating officials, encourages counseling early in the rating period and enhances the effectiveness of organizations by relating performance to organizational missions. The senior rater concept increases the role of the most senior rating official from a purely administrative role to include a critical evaluation of the rated officer's potential. The senior rater profile provides a comparison of a specific rating and a senior rater's normal rating tendency (e.g., easy or hard) by tracking the rating history of the senior rater and making it visible to selection boards and DA managers.

THE "NEW" ARMY OFFICER EVALUATION REPORTING SYSTEM

Richard D. Doorley
U.S. Army Military Personnel Center

The U.S. Army transitioned to a new Officer Evaluation Reporting System (OERS) during the latter part of 1979. The development of the new system took over six years during which several thousand officers from the field participated in one way or another in the developmental process, and, to a large extent, determined the makeup of the new system. The process included: an Army-wide field test conducted in 110 Active Army, National Guard and Reserve organizations, an Army-wide survey, extensive involvement with selection boards and career managers; a review of the regulation and supporting policies with the major commands; and a review of the performance evaluation systems of sister services, government, industry, academia, and many allied military services.

The performance evaluation literature reveals many commonalties. The stated purpose in most organizations is to provide information for making traditional management decisions: promotion, assignments, awards, training and demotion. The overall trend is away from evaluating personality traits to management by objectives techniques with employee participation and feedback requirements. All evaluation systems suffer from varying degrees of inflation which appears to be positively related to the importance placed on the ratings. Few organizations exercise any centralized control over raters but most have a requirement for additional evaluators and/or reviewers to insure fairness in the evaluation process.

Past Army Systems

Prior to the 1920's Army officer management decisions were handled at unit level based primarily on personal knowledge. After World War I officer personnel management was centralized and a standardized form developed, Form 67, which was adopted in 1922. This form, with minor revisions, was used through World War II when massive reductions put extreme pressure on the Officer Evaluation System (OES) and resulted in major changes. Since then there have been eight editions of the OER caused primarily by turbulence in officer personnel management and the concommitant pressures to inflate ratings, especially with an up or out promotion system.

Numerous evaluation techniques and methods have been employed on past forms: narratives, rating scales, forced choice items, forced distribution scales, cumulative numerical scores, and "closed" or secret reports. The success of the various techniques was dependent upon acceptance by the officer corps, how well they were designed and regulated and in some cases when they were used.

In 1973 shortly after the implementation of DA Form 67-7, serious doubts were raised concerning its effectiveness. It inflated rapidly as published benchmark scores became floors rather than Army-wide averages. Lack of confidence and support from the officer corps and leadership led further to its demise; consequently, the development of a follow-on OER began early in 1973.

The Officer Evaluation Reporting System (OERS) is a subsystem of the Officer Evaluation System. It includes procedures for organizational evaluation chain assessment of an officer's performance and an estimation of potential for future service based on the manner of that performance. The major function of the OERS is to provide information from the organizational rating chain to Department of the Army (DA) for officer personnel decisions. The other two functions are to encourage the professional development of the officer corps and to enhance mission accomplishment. To support these functions the new OERS incorporates several new features which have not been included in previous officer evaluation reporting systems.

The New Officer Evaluation Report

During the field test, the increased involvement of the rated officer in the evaluation process was strongly endorsed. Rated officer input in the form of developing the duty description and performance objectives was viewed as an ideal technique for increasing two way communications between the rater and rated officer, especially in terms of developing and clarifying the elements of the rated officer's performance. Rated officers perceived an increased awareness of the specific nature of the job as well as an increased opportunity to influence decisions on mission accomplishment. Rating officials gained valuable insight into the status of the organization along with additional information upon which to base an accurate evaluation. Now for the first time the rated officer is included in the rating process with the use of DA Form 67-8-1, OER Support Form. Participation by the rated officer addresses all three functions of the evaluation system. It provides additional information to the rating chain from the rated officer's point of view, encourages two-way communication and professional development, and increases the effectiveness of organizations by focusing performance more directly on the mission.

The front side of the Support Form (Figure 1) contains rated officer and rating chain identification and provides for the rated officer's description of his duties, major objectives and contributions during the rating period. The reverse side (Figure 2) provides for rater and intermediate comments on the rated officer's input, as well as instructions for completing the form. The Support Form is for rated officer and rating officials use and is not forwarded to DA. At the beginning of the rating period it is used as a guide for discussion between the rater and rated officer about the rated officer's duties, responsibilities and performance objectives for the period. During the evaluation period, it acts to guide the performance of the rated officer and the counseling and coaching by the rater. At the end of the rating period, it gives the rated officer an opportunity to provide the rating chain information about his performance from his point of view.

At the beginning of the rating period, the rated officer receives a copy of the OER Support Form which is blank except for the name, grade and position of his rater and the positions of the remaining members of the rating chain. Within the first thirty days of the rating period, the rated officer and rater are required to discuss the specific nature of the rated officer's duties (his duty description) and the focus or direction of his performance (his major performance objectives). These objectives can be developed in several ways. They can be set by the rater, suggested by the rated officer (which may be appropriate when the rater is new or recently assigned) or may be jointly developed. Furthermore, this task can be accomplished formally or informally depending on the situation, style or personality of the rater. The important thing is that the discussion takes place and the rated officer gets started in the proper direction.

During the rating period, the rater and rated officer should update the duty description and performance objectives to reflect shifts in emphasis, additional missions, and other changes that are likely to occur. This updating process presents the rater with an ideal opportunity to coach or counsel the rated officer and provide additional guidance as well as the benefit of his knowledge and experience. Concurrently, the rated officer is afforded an opportunity to provide relevant comments, perceptions and suggestions concerning his performance and future direction.

At the end of the rating period, the rated officer receives a Support Form with identification data filled out to include the names of the senior rater and the intermediate rater if one has been included in the rating chain. The rated officer is required to complete his duty description, major performance objectives, and significant contributions--and submit the Support Form to his rating chain. If the rated officer has kept his information updated and has continued to communicate with his rater, he is in the best position to provide objective comments concerning his performance and avoid unrealistic remarks.

The Support Form institutionalizes a procedure that has been a part of effective military leadership for generations - that of clarifying performance expectations for subordinates and, in the process, assisting the subordinate in his professional development. Properly used, the Support Form is beneficial to the rated officer and rating officials not only in attaining a more complete and valid evaluation, but also in improving performance.

Senior Rater Concept

One of the fundamental concepts of the new system is to increase the role of more senior officers in the rating chain. The intent is to fix a critical evaluation responsibility on the individual in each rating chain who is at least one level removed from the immediate supervision of the rated officer, and yet close enough to the rated officer to be aware of the organizational circumstances surrounding his performance. By being one level removed from direct supervision, the senior rater has additional experience to judge performance from a broader perspective, has a wider range of officers to consider and compare, and is more likely to weigh organizational requirements and actual performance results more heavily than personal relationships and personality. He evaluates the potential of the rated officer in comparison with a sample population of 100 officers of the same grade. This is accomplished by placing an X in the appropriate box in the left hand portion of the distribution (See Figure 3). While there are no quotas or rigid requirements to spread the rated officers across all of the boxes, logic imposes its own constraints on the number of officers who can be placed in any particular block. For example, it is extremely unlikely that all of the officers rated by a senior rater will always be one in a hundred. Therefore, the senior rater who consistently places all his officers in the top box is distorting the system and is simply not supplying DA with credible rating information.

When the OER arrives at Department of the Army Military Personnel Center (MILPERCEN), the rating history or profile of the senior rater is placed in the boxes in the right hand portion of the distribution. This history shows exactly how the senior rater placed all officers he evaluated of the same grade as the rated officer up to that time. For example, a senior rater evaluates a major by placing him in the second box on the left. When the report is accepted by Department of the Army MILPERCEN as correct, the number of majors senior rated by that senior rater up to and including that report will be placed in the boxes on the right. This provides selection boards with a comparison of the senior rater's general tendency and how he rated that particular officer, thereby addressing that age old problem of hard versus easy raters (See example at Figure 4). In addition, this profile will offer some protection to officers rated early in the new system in the event of general inflation later on in the life of the system. Even after several years, a selection board will still be able to see exactly how the senior rater was evaluating officers the day he rated the individual officer. After evaluating potential, the senior rater makes any appropriate comments on performance, potential or anything associated with his review of the entire form. At this point the senior rater forwards the OER through the servicing Military Personnel Office to Department of the Army MILPERCEN and returns the Support Form to the rated officer.

The DA Form 67-8-2, Senior Rater Profile Report (Figure 5) is used by Department of the Army to track and maintain a record of the rating history of the senior rater. This rating history is expressed in terms of the number of reports rendered and the number of different officers evaluated. It is produced annually for all senior raters who have senior rated at least five different officers. One copy of the report is made available to the senior rater and one copy is placed in the senior rater's Official Military Personnel File.

The purpose of the senior rater profile report is to remind all senior raters of their responsibility to supply credible evaluative information to Department of the Army. This is one of a senior rater's most important responsibilities because of its impact on the future leadership of the Army. Therefore, because it is an important responsibility, it is an element of performance and is placed in the senior rater's performance fiche next to his other performance documents. Thus, the senior rater profile report is an indication to selection boards as to the degree to which a senior rater accepts his evaluation responsibilities.

Discussion

The senior rater concept was instituted to help dampen inflation by taking the pressure off the subordinate's immediate supervisor and giving the critical potential evaluation to the more senior individual in the rating chain who has a broader organizational perspective and theoretically a more objective basis for evaluation. Some common causes of inflated ratings - loyalty to the individual subordinate, team cohesion, perceived superiority of subordinates, the perception of high levels of responsibility and communication difficulties - affect the senior rater to a lesser degree than the immediate supervisor; however, the Army's up or out promotion policy still remains a pervasive pressure even for the senior rater. For the first time in history some type of control is maintained over ratings given by senior raters. Selection boards are able to see how senior raters shoulder and accept their rating responsibilities via the Senior Rater Profile Report which is generated annually for senior raters. Although the "new" OER System has been in effect for almost two years with over 200,000 reports received, it is still too early to make a conclusive assessment of how well it is accomplishing the objectives for which it was designed; however, preliminary feedback from the field and selection boards is very positive. The use of the OER Support Form appears to be making a significant contribution not only to the evaluation process, but also the goal of better performance. Field reaction has been extremely positive, with several benefits consistently reported: increased awareness by rated officers of their responsibilities, a closer alignment of performance to organizational missions, an opportunity for rated officers to remind rating officials of what was accomplished during the rating period, and the availability of specific information to rating officials, making preparation of the OER easier than it has been in the past.

Feedback from selection boards to date (40 boards, more than 300 members) has provided insight into both positive and problematic areas of the new system. Responses to a selection board questionnaire indicate a healthy balance exists between the rater and senior rater portions of the OER reflecting the importance of both to the evaluation process. (See Figure 6 and 7 for complete OER, DA Form 67-8). There does not appear to be undue focus on the senior rater portion of the OER. Early fears from the field that a top block evaluation was needed for promotion have been dispelled by selection board follow-up studies. In almost a full year's cycle of selection board deliberations, more than half of the selectees had less than top box evaluations with a range of 1-5 box ratings, even when selection rates were as low as 5%.

Senior raters Army-wide are shouldering their responsibilities very well. The vast majority appear to be spreading their effective, successful officers over at least the top four boxes. However, it must be remembered that the worth of a senior rater evaluation is not based on what all senior raters throughout the Army did, but rather, on the comparison of the box check with the individual senior rater's general rating tendency or profile, as amplified and explained by his remarks concerning the rated officer.

OFFICER EVALUATION REPORT SUPPORT FORM

For use of this form, see AR 623-105; proponent agency is US Army Military Personnel Center

Read Privacy Act Statement and Instructions on Reverse before Completing this form.

PART I – RATED OFFICER IDENTIFICATION

NAME OF RATED OFFICER (Last, First, MI)	GRADE	PRINCIPAL DUTY TITLE	ORGANIZATION

PART II – RATING CHAIN – YOUR RATING CHAIN FOR THE EVALUATION PERIOD IS:

RATER	NAME	GRADE	POSITION
INTERMEDIATE RATER	NAME	GRADE	POSITION
SENIOR RATER	NAME	GRADE	POSITION

RATED OFFICER'S SPECIALTIES/MOS ______ DUTY SSI/MOS ______

PART III – RATED OFFICER (Complete a, b and c below for this rating period)

a. STATE YOUR SIGNIFICANT DUTIES AND RESPONSIBILITIES

b. INDICATE YOUR MAJOR PERFORMANCE OBJECTIVES

c. LIST YOUR SIGNIFICANT CONTRIBUTIONS

______
(Signature and Date)

DA FORM SEP 79 67 – 8 – 1

FIGURE 1. FRONT SIDE OF OER SUPPORT FORM

PART IV – RATER AND/OR INTERMEDIATE RATER *(Review and comment on Part III a, b, and c above. Insure remarks are consistent with your performance and potential evaluation on DA Form 67-8.)*

a. RATER COMMENTS *(Optional)*

SIGNATURE AND DATE *(Mandatory)*

b. INTERMEDIATE RATER COMMENTS *(Optional)*

SIGNATURE AND DATE *(Mandatory)*

DATA REQUIRED BY THE PRIVACY ACT OF 1974 (5 U.S.C. 552a)

1. AUTHORITY: Sec 301 Title 5 USC; Sec 3012 Title 10 USC.

2. PURPOSE: DA Form 67–8, Officer Evaluation Report, serves as the primary source of information for officer personnel management decisions. DA Form 67–8–1, Officer Evaluation Support Form, serves as a guide for the rated officer's performance, development of the rated officer, enhances the accomplishment of the organization mission, and provides additional performance information to the rating chain.

3. ROUTINE USE: DA Form 67–8 will be maintained in the rated officer's official military Personnel File (OMPF) and Career Management Individual File (CMIF). A copy will be provided to the rated officer either directly or sent to the forwarding address shown in Part I, DA Form 67–8. DA Form 67–8–1 is for organizational use only and will be returned to the rated officer after review by the rating chain.

4. DISCLOSURE: Disclosure of the rated officer's SSAN (Part I, DA Form 67–8) is voluntary. However, failure to verify the SSAN may result in a delayed or erroneous processing of the officer's OER. Disclosure of the information in Part IIIc, DA Form 67–8–1 is voluntary. However, failure to provide the information requested will result in an evaluation of the rated officer without the benefits of that officer's comments. Should the rated officer use the Privacy Act as a basis not to provide the information requested in Part IIIc, the Support Form will contain the rated officer's statement to that effect and be forwarded through the rating chain in accordance with AR 623–105.

INSTRUCTIONS

PART I: Identification – Self explanatory.

PART II: Rating Chain – The personnel officer or appropriate administrative office will fill in information based on the commander's designated rating scheme.

PART IIIa: Rated Officer Significant Duties and Responsibilities – State the normal requirements met in your specific position as well as any important additional duties. Address the type of work required, rather than frequently changing specific tasks.

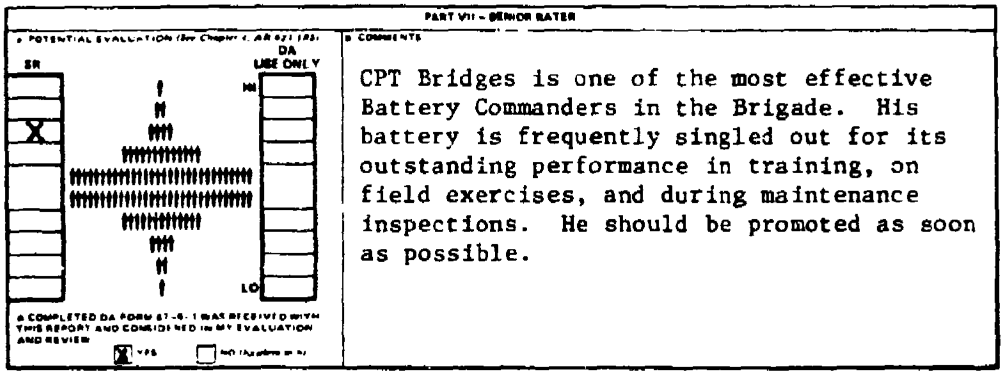
PART IIIb: Rated Officer Major Performance Objectives – List the most important tasks, priorities, and major areas of concern and responsibility assigned. This is an explanation of how you set out to accomplish the duties described in IIIa. Ideally these are planned goals that you will work toward in an effort to make a contribution to the accomplishment of the organization mission, however, they may be in reaction to unpredictable changes. The objectives come from the following four categories:

> **ROUTINE** – Objectives that address the repetitive a[nd co]mmonplace duties that must be carried out. These are duties that will produce less visible results [but] will have serious consequences if not properly executed.
>
> **PROBLEM SOLVING** – Objectives that provide [for de]aling with problem situations. The objective should plan for or address potential problems so t[hat tim]e is available to deal with them without disrupting other objectives.
>
> **INNOVATIVE** – Objectives that create new or improved me[th]ods of operation in the organization.
>
> **PERSONAL DEVELOPMENT** – Objectives that further professional growth of an individual or his/her subordinates.

PART IIIc: Rated Officer Significant Contributions – Describe the most significant contributions you made during the rating period. These may have been in support of the objectives established or may highlight other accomplishments that you feel are important.

PART IV: Rater and/or Intermediate Rater Review and Comment – Insure any remarks are consistent with your performance and potential evaluation on DA Form 67–8. Signature does not show concurrence with Part III but indicates that you have reviewed the rated officer's portion of the form.

FIGURE 2. REVERSE SIDE OF OER SUPPORT FORM



PART VII - SENIOR RATER

SR

DA USE ONLY

HI

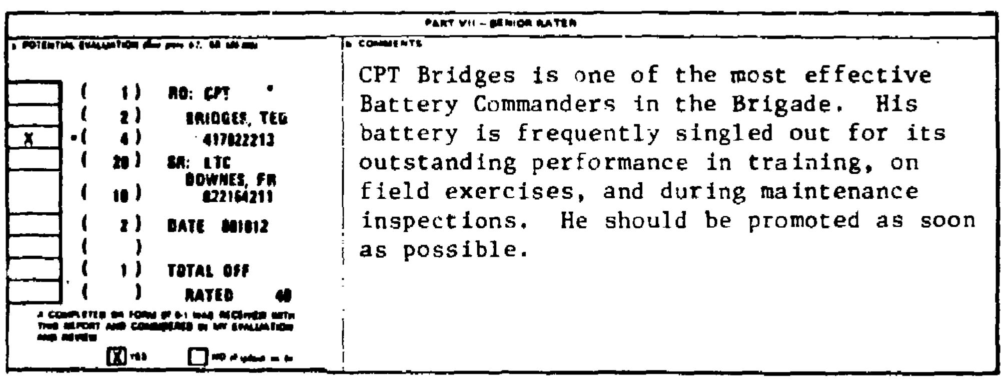
X

LO

YES

CPT Bridges is one of the most effective Battery Commanders in the Brigade. His battery is frequently singled out for its outstanding performance in training, on field exercises, and during maintenance inspections. He should be promoted as soon as possible.

FIGURE 3. SENIOR RATER POTENTIAL EVALUATION



PART VII - SENIOR RATER

(1) RO: CPT
(2) BRIDGES, TEG
X (4) 417022213
(20) SR: LTC
BOWNES, FR
(10) 022164211
(2) DATE 001012
( )
(1) TOTAL OFF
( ) RATED 40

YES

CPT Bridges is one of the most effective Battery Commanders in the Brigade. His battery is frequently singled out for its outstanding performance in training, on field exercises, and during maintenance inspections. He should be promoted as soon as possible.

FIGURE 4. SENIOR RATER POTENTIAL EVALUATION WITH DA GENERATED LABEL AFFIXED DEPICTING CUMULATIVE SENIOR RATER PROFILE FOR CPTs.

SENIOR RATER PROFILE REPORT
OFFICER EVALUATION REPORTING SYSTEM
For use of this form, see AR 623-105; proponent agency is US Army Military Personnel Center.

PART I – ADMINISTRATIVE DATA

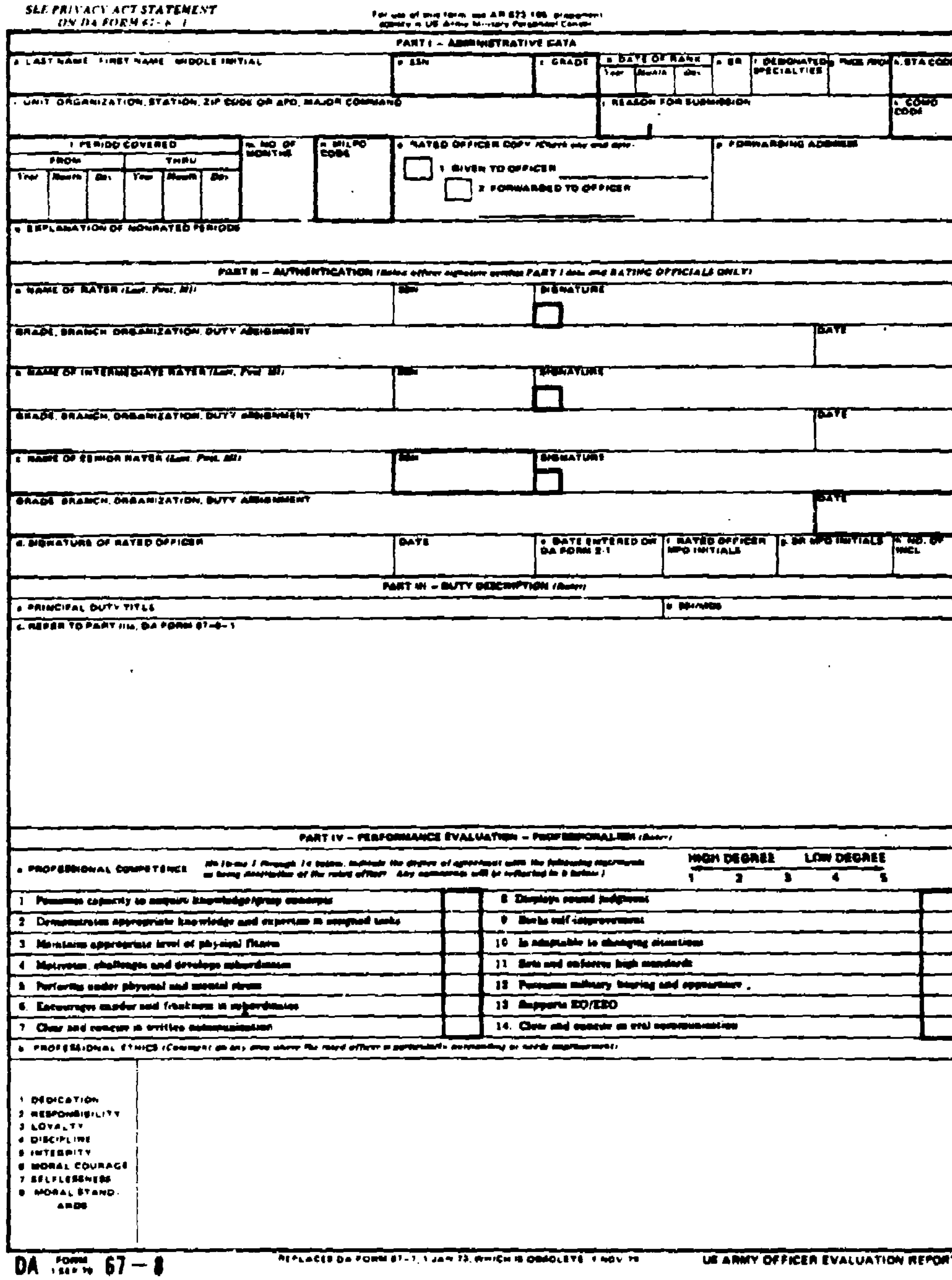
a. NAME *(Last, First, MI)*	b. SSN	c. GRADE	d. DATE OF REPORT
DOE, JOHN J	000-11-2222	BG	25 SEP 81

PART II – SENIOR RATER PROFILE

MG	BG	COL	LTC	MAJ	CPT	2LT / 1LT	CW3 / CW4	WO1 / CW2	TOTAL RATINGS	HIGHEST
			2	4	1				7	
			4	4	5				13	
			6	8	5				19	
			4	3	2				9	
			2	2	1				5	
			0	1	1				2	
			1	0	0				1	
			0	0	0				0	
			0	0	0				0	
									56	LOWEST
			18	20	13				TOTAL OFFICERS 51	

DA FORM 1 SEP 79 67 – 8 – 2

FIGURE 5. EXAMPLE OF ANNUAL SENIOR RATER PROFILE REPORT


SEE PRIVACY ACT STATEMENT ON DA FORM 67-8-1

PART I – ADMINISTRATIVE DATA

PART II – AUTHENTICATION

PART III – DUTY DESCRIPTION

PART IV – PERFORMANCE EVALUATION – PROFESSIONALISM

HIGH DEGREE LOW DEGREE

1 2 3 4 5

1 Possesses capacity to acquire knowledge/grasp concepts

2 Demonstrates appropriate knowledge and expertise in assigned tasks

3 Maintains appropriate level of physical fitness

4 Motivates, challenges and develops subordinates

5 Performs under physical and mental stress

6 Encourages candor and frankness in subordinates

7 Clear and concise in written communication

8 Displays sound judgment

9 Seeks self improvement

10 Is adaptable to changing situations

11 Sets and enforces high standards

12 Possesses military bearing and appearance

13 Supports EO/EEO

14 Clear and concise in oral communication

1 DEDICATION
2 RESPONSIBILITY
3 LOYALTY
4 DISCIPLINE
5 INTEGRITY
6 MORAL COURAGE
7 SELFLESSNESS
8 MORAL STANDARDS

DA FORM 1 SEP 79 67-8

US ARMY OFFICER EVALUATION REPORT

FIGURE 6. FRONT SIDE OF OFFICER EVALUATION REPORT

PERIOD COVERED

PART V – PERFORMANCE AND POTENTIAL EVALUATION (Rater)

a. RATED OFFICER'S NAME ______ SSN ______

RATED OFFICER IS ASSIGNED IN ONE OF HIS/HER DESIGNATED SPECIALTIES/MOS ☐ YES ☐ NO

b. PERFORMANCE DURING THIS RATING PERIOD (REFER TO PART III, DA FORM 67–8 AND PART III a, b, AND c, DA FORM 67–8–1)

☐ ALWAYS EXCEEDED REQUIREMENTS ☐ USUALLY EXCEEDED REQUIREMENTS ☐ MET REQUIREMENTS ☐ OFTEN FAILED REQUIREMENTS ☐ USUALLY FAILED REQUIREMENTS

c. COMMENT ON SPECIFIC ASPECTS OF THE PERFORMANCE (REFER TO PART III, DA FORM 67–8 AND PART III a, b, AND c, DA FORM 67–8–1. DO NOT USE FOR COMMENTS ON POTENTIAL)

d. THIS OFFICER'S POTENTIAL FOR PROMOTION TO THE NEXT HIGHER GRADE IS

☐ PROMOTE AHEAD OF CONTEMPORARIES ☐ PROMOTE WITH CONTEMPORARIES ☐ DO NOT PROMOTE ☐ OTHER (Explain below)

e. COMMENT ON POTENTIAL

PART VI – INTERMEDIATE RATER

a. COMMENTS

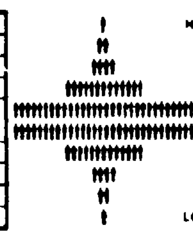
PART VII – SENIOR RATER

a. POTENTIAL EVALUATION (See Chapter 4, AR 623-105)

SR

DA USE ONLY

HI



LO

A COMPLETED DA FORM 67–8–1 WAS RECEIVED WITH THIS REPORT AND CONSIDERED IN MY EVALUATION AND REVIEW ☐ YES ☐ NO (Explain in b)

b. COMMENTS

FIGURE 7. REVERSE SIDE OF OFFICER EVALUATION REPORT

AD P001310

NROTC Joins the BIG Test Movement

ANDREW N. DOW, Ed. D
Director of Naval Science Evaluation
CNET, Pensacola, Florida

SUMMARY

Traditionally, CNET has allowed the local NROTC units autonomy in matters of testing and grading. The high level of students and institutions assured adequate quality. Recently, a few NROTC graduates were having difficulty completing the post-commissioning schools, especially Surface Warfare Officers' School (SWOS). If NROTC is to anticipate who will have SWOS problems, a comprehensive instrument to evaluate each midshipman's retention and comprehension of the first three years work is needed.

The exam was developed using the philosophy and techniques that have evolved with the Navy enlisted advancement examination program. Because the use of the exam and the population to be tested are different from that of the advancement program, a number of changes became mandatory. The three hour exam has 150 items in five subject matter areas. Total standard scores and section stanines were calculated; item analysis data were extracted. Persons who fell into the lower stanine on any section were marked for remediation. Units tailored their remedial work to those that needed it. No unit statistics were computed as there is no desire to set up inter-unit rivalries.

For the 1981 exam and future exams, new software developed by TAEG of Orlando will pinpoint specific difficulties and calculate section standard scores in addition to providing the data produced in 1980. This will enable individuals to zero-in on their deficiencies. Semi-automated production is another goal.

The Naval Reserve Officers' Training Corps (NROTC) consists of fifty five (55) units located on the campuses of major universities and colleges. At some of these units, students from non-host institutions are "cross-town" enrolled; thus the midshipmen of NROTC are pursuing degrees at approximately 100 colleges and universities. The Naval Science training is spread over the four years of the undergraduate career; it consists of eight courses and three summer cruises. Each course is based upon a Chief of Naval Education and Training (CNET) approved curriculum, and is supervised by CNET with the assistance of a course-coordinating NROTC unit. One officer instructor attached to the course-coordinating unit is designated as the point-of-contact which makes him the active course coordinator. Specific details of course presentation, testing, and grading are local matters under the purview of the unit commanding officer who is the university's Professor of Naval Science.

The four year Navy-option program, which includes some special requirements outside the department of Naval Science, leads to a commission as ENSIGN. Midshipmen who are enrolled as scholarship students are commissioned into the regular Navy; non-scholarship (college program) students are commissioned into the Naval Reserve. Scholarship students receive $100 per month plus books and tuition. Newly commissioned naval officers, regardless of source, generally attend a post-accession school--flight, supply corps, submarine, surface warfare, etc. There is also a Marine option with some different courses; this leads to a commission as a second lieutenant in the Marine Corps or the Marine reserve.

Several years ago, it was observed that a higher percentage of NROTC graduates were being attrited from the Surface Warfare Officers' School Basic Course, than were officers from other sources (U.S. Naval Academy and OCS). The Naval Personnel Research and Development Center, San Diego was tasked with investigating this matter and reporting their findings. This report (Crawford) pointed the finger at the NROTC program on a superficial level, but they had not investigated causes nor specific sources of the problem. Informal tallies of the specific NROTC units of the attriting officers tended to show that some campuses were consistent sources of attrited officers, and that there were a few that were intermittent sources of attrites from the Surface Warfare Officers' School. Independently, the Navy Inspector-General's 1978 inspection team suggested that CNET develop end-of-course tests for all Naval Science courses. CNET chose instead to develop a corps-wide comprehensive examination to be administered to all first-class Navy option midshipmen (seniors) early in the academic year. The dual purpose of this examination is to encourage quality instruction and to diagnose deficiencies so that they may be remedied during the final year; it is not a hurdle to be cleared before commissioning. Headquarters is not usurping the right of the local unit commanding officer to decide who should be commissioned and who not. This exam, hopefully, will assist the staffs of the local units in identifying those midshipmen that have forgotten some of the crucial material they previously learned. While there are some questions that check on direct, rote knowledge of nomenclature, the whole exam is slanted toward the broader aspects of information and the application of interacting items of knowledge--it is aimed at macro-objectives rather than micro-objectives.

The broad philosophy of the examination is similar to that of the successful examinations used in the Enlisted Advancement program. Both the NROTC and the Navy Enlisted Advancement program assume that everyone who takes the exam is qualified to undertake the next pay-grade up the ladder. The purpose of both examination programs is to find out who is better qualified; at this point, the two programs go their own ways. NROTC tutors those who need it; the Advancement system sets the less qualified aside to mature and study on their own so that they may do better next time. Both programs use examinations that are designed to distinguish between individuals over a wide range of performance. In other words, these exams spread the members of the upper and lower quarters with approximately equal validity.

Strictly speaking, the NROTC COMPrehensive could have been cast as a mastery type exam. But, there are a number of reasons why a mastery-type exam would be less than satisfactory for this purpose. First, a realistic mastery-type exam that includes only materials which MUST be mastered will appear as an easy waste of time to the top three-quarters of an elite population like the seniors of the NROTC midshipmen, and those who do poorly are labeled "dummies." Second, using a truncated measure such as a mastery test is analogous to measuring the height of American men with a ruler that is 65 inches long; the distribution of the scores will be similarly distributed in both of these cases. Of course, there may be a situation in which a specific project would need to know how many men fell into each inch category of height up to and including 65 inches, and how many were over that figure. The current needs of the NROTC program are such that a truncated information exam might be satisfactory from some viewpoints, but it would leave many questions unanswered. Third, mastery-type exams tend to dwell upon specific minutiae, or the microobjectives of an instructional program or of a set of specific "competencies." While the specific micro-objectives of an instructional program are the necessary steps in the development of an understanding of the macro-objective of a series of lessons, it is not necessary that the person who understands and comprehends the macro-objective be able to recall specific bits of micro-objective related information. Fourth, it is conceivable that the NROTC COMPrehensive Examination might, in the future, serve as the basis for recognizing those midshipmen who have achieved a high degree of comprehension of the professional content of the officer training program. The truncated measure is completely useless for this purpose. Fifth, there are grounds to argue that a mastery-type exam yields a score that is a tabulation of the number of "digital" successes achieved, and is not a measure of a continuous phenomenon. The conventional objective examination, sometimes referred to as "norm-referenced", yields scores that fall along a true continuum. Sixth, most consumers of test scores have some understanding of, and expect, statistical analyses that utilize the parametric statistics developed by Galton and Pearson. These statistics are meaningless unless based upon normally distributed continuous data (Treloar). Meaningful statistics cannot be derived from the non-normally distributed discontinuous data that tallies the specific successes achieved on a mastery type test.

The content of the NROTC COMPrehensive exam is based both on the curriculum and upon the broad competencies expected of a junior naval officer. Its score conversion is based upon the performance of a population of known ability that has been instructed according to a standard set of curricula. Thus the NROTC COMPrehensive is curriculum-and-competency based and population referenced. It is not norm-referencd, NROTC is not developing a set of norms against which to compare others. The population sets the performance level, thus it is not criterion referenced. It does yield continuous data that distribute themselves rather normally along a continuum. Therefore, valid standard deviations, standard scores, correlations, etc., can be computed, and used meaningfully.

The techniques used to develop the exams for the Naval Enlisted Advancement system were adapted to the special needs of the NROTC examination. The advancement system uses fleet-experienced senior petty officers to produce the raw materials from which its highly successful examinations are built. NROTC, analogously, uses line officers with recent experience in the role of OOD, in engineer duty, and in the weapons field to write questions for the pool from which the exams are constructed. In contrast to the petty officers who are assigned to examination work for a tour of two or more years, the test writing is a special duty of the line officers who instruct in the NROTC units. Periodically, these officer instructors are invited to prepare and submit one (more if they choose) test question for each lesson in the courses they are responsible for. There is no point in delving into the actual nitty-gritty of building and publishing the exams, suffice it to say that NROTC's procedures are derived from those used with the advancement system. In the semi-final stages, the exam as a whole is reviewed and the answers checked by several line officers.

Specifically, why is this a BIG examination? First, it is big because of its scope--it covers the high points of three years of Naval Science courses. These courses are the introduction to the Navy and its traditions, Naval Ships' Engineering Systems, Naval Ships' Weapons Systems, Navigation, Ship Operations, and Seapower and Naval History. Secondly, it is a three hour exam with 150 multiple choice questions. Lastly, this is a BIG examination because it is used nationwide.

The examination consists of five sections with thirty questions in each; the introduction to the Navy and the History of Sea Power are presently merged into a single section. Each of the other four courses is examined by a full section of thirty items. Whenever practical, questions are problem oriented. In the section devoted to ship operations there are several problems to be worked out on paper maneuvering boards (relative motion plots). At least two, and in some cases as many as four, questions are asked about the results of each maneuvering board problem. In the 1980 examination each section was a discrete block of questions; but in the 1981 examination each section is spread from the beginning to the end in spiral fashion. This is part of the attempt to make the NROTC examination as realistic as possible--in real life, problems do not appear in neat well-classified packages. Improved software makes it possible to derive section scores and section related item analysis data from spiraled sections.

The 1980 NROTC examination papers were processed using the hardware and software used for some of the examinations in the enlisted advancement system. The system yielded a total raw score, section raw scores, a Navy standard score conversion for the total score, and section stanines for each participant. A graphic frequency distribution portrayed the overall performance of all participants; it was augmented by parametric descriptive statistics. The system also yielded item analysis data that included the total response to each alternative response to the question, the overall difficulty of each item, and each item's intra-section discriminatory power. Score lists were prepared for each individual NROTC unit listing the score data for each of the midshipmen attached to the specific unit. Any participant that scored in the lowest stanine on any section was tagged for remediation in the specific field. This cut-off point approximates the Navy's time-honored passing score of -1.5 sigmas (standard score of 35). Unit averages and other unit-specific data were not computed; the purpose of the exam is the improvement of the individual rather than inter-unit competition.

The Chief of Naval Education and Training (CNET) charged each local commanding officer (Professor of Naval Science) with the responsibility of devising and implementing a program fitted to the needs of the persons in his command that needed remediation. CNET also requested that the nature and results of the remedial program be reported. As little guidance came from the command, there was a wide variety of programs. These programs cannot be evaluated at present because the participants are members of the class of '81, and their success as Naval officers and post-accession trainees is unknown.

The 1981 NROTC COMPrehensive Examination will be processed using software developed by the Training Analysis and Evaluation Group at Orlando (TAEG). In addition to the data supplied by the advancement system software, the new software will produce a list of individuals (by unit) that shows which items each missed. This will reduce the time and effort expended in remedial work, and should make that which is done more effective. There is no reason for a midshipman to repeat the entire semester's course when he is weak in about a quarter of the course.

As the bank of useful examination questions grows, the process of putting them into a magnetically recorded bank will begin. Materials from such a bank could be used to produce the exams in a semi-automated fashion. The person who puts the new exam together will be able to call for print-outs of items in a specific area, then select and/or revise them. After the tentative selection has been made, the word processor will print out the selected revised items. After some hand-massaging of the tentative exam, the word processor will produce the originals of the final product.

Another proposed future development lies in the area of switching from "content-oriented" questions to "process-oriented problems." Some of the present questions are more concerned with "how" than with "what" or "why," but in the future, more of them will be "how" items supplemented with some essential "why" items. There may even be room for a few "what" items. The general validity of the process-oriented question, the "how" item, is being established by the College Outcomes Measurement Project of the American College Testing Program (COMP).

As for the effectiveness of NROTC's BIG examination—as stated earlier, it is too new to evaluate. The internal descriptive statistics are acceptable for a new untried test. TAEG will perform a predictive validity study when sufficient data and criteria are available. This and other studies will be reported in the future when they are available.

References

COMP: College Outcome Measures Project Summary Report of Research and Development 1976-1980. Iowa City: American College Testing Program, 1980.

Crawford, A. and Van Beenan, S. Evaluation of Officer Accessions at Surface Warfare Officer School NRPDC TR 80-24 (AD-A085028).

Treloar, Alan E. Elements of Statistical Reasoning. New York: John Wiley & Sons, Inc., 1939.

AD P001311

WORK CHARACTERISTICS -- A TASK-BASED, BENCHMARK APPROACH

WALTER E. DRISKILL, Ph.D.

MR. JAMES B. KEETH

MAJOR FRANK C. GENTNER

OCCUPATIONAL ANALYSIS PROGRAM
USAF OCCUPATIONAL MEASUREMENT CENTER
AIR TRAINING COMMAND
RANDOLPH AFB, TEXAS 78150

Most occupations involve a number of work characteristics and aptitudes. Present AQE/ASVAB categories or "areas of enlistment" do not fully describe the nature of the work being performed within Air Force occupational specialties. Selection of personnel on the basis of AQE/ASVAB categories frequently leads, as a consequence, to personnel expecting work to be predominantly a particular kind when, in fact, other work requirements are paramount. Because of this problem, USAFOMC developed a task-by-task look at USAF occupations using a benchmark scale to categorize the nature of work being performed within each Air Force specialty. Teams of subject matter experts were presented lists of commonly performed tasks and assisted occupational analysts in categorizing the tasks according to the benchmark scales. Most occupations were found to consist of tasks fitting a predominant category, with several other categories present as well. Some previously "Electronics" specialties, for example, were found to be primarily "Electrical," with a high component of mechanical skills. This paper presents the rationale, methods, and preliminary results of the USAFOMC effort.

That the aptitude indices of the Armed Services Vocational Aptitude Battery (ASVAB) are not perfectly descriptive of the nature of the occupations for which they are used for selection is not a startling revelation. That, because of long custom, they have taken on a far greater meaning than they are entitled is no surprise. But, as a result of the broad and indefinite meaning attached to aptitude scores, several problems, as well as misunderstanding, in the recruitment, classification, and utilization of personnel arise. One effort to provide more insight into occupations is a study initiated by the Air Force Occupational Analysis Program to describe the components of the enlisted specialties. The study involves a task-based, benchmark methodology.

Since 1948 (Massey and Creager, 1956) the Air Force has used some form of classification test as a selection device. These classification tests are validated against school success (Mullins, et al, 1981), a practice justified at least by the absence of stable and relevant performance criteria. In addition, one might speculate about the value of predictive or concurrent validation efforts, because of the effects of experience.

Even when validated against school success, the relationship of aptitude is unclear. The General Aptitude Index, for example, often is as efficient a predictor of success in training as another index that is in practice used to select for an Air Force specialty (Vitola, et al, 1973). The bottom line is that we know little, but make many assumptions, about aptitude and job performance.

A kind of magic is attributed to the aptitude indices. Operators, for example, contend that they need personnel with a higher aptitude than the minimum used in the classification of a specialty. No one can argue successfully against the desirability of higher aptitude personnel for most specialties, since there is a dearth of information about the relationship of aptitude to job performance. On a more specific level, however, the operators may need more experienced personnel rather than higher aptitude workers. The problem here may be one of articulation of needs.

The lack of congruence between aptitude indices and the components of the jobs for which they are used to select also creates a greater problem. Personnel are counseled during recruiting and initial classification into the Air Force on the kind of Air Force specialty for which they are qualified. They are often recruited for an "electronic" specialty because they have acceptable Electronic Aptitude Index (EAI) scores. Sometimes, however, the "electronic" specialty is not, in fact, an electronic occupation even though the EAI is the selection criterion. Dissonance results, as in a recent case where an airman with an EAI score of 95 was assigned to a missile maintenance specialty for which EAI score is the selection criterion. This airman believed he was getting into an electronic field, but he observed very soon that in his job, he encountered nothing electronic. He complained. As a result, we made a task-by-bask analysis of the specialty. Not only did we find that he was correct about his job but that no electronic task -- or a task requiring electronic knowledge -- existed in the specialty.

This experience, along with others, led the Air Force Manpower and Personnel Center Classification Branch to request the U.S. Air Force Occupational Analysis Program to study all enlisted specialties to identify the components of each. The remainder of this paper describes this study.

APPROACH

The initial phase of the project involved an extensive review of the literature to get a good historical perspective of the aptitude research conducted by the USAF Human Resources Laboratory (AFHRL), primarily relating to development of the Airman Qualification Examination (AQE) and Armed Services Vocational Aptitude Battery (ASVAB) and derivation of their categories or subcategories. In addition, discussions were held with AFHRL researchers familiar with this area. Once review of the literature was completed, definitions were tentatively derived for each of the present four AQE/ASVAB categories - Administrative, Electronic, Mechanical, and General - and subskills or components involved were related to each area. This phase of the project turned out to be much more difficult than at first imagined, since in reviewing the literature, no one "definitive" definition of the four areas could be found, and in some cases, no agreement was found as to what specific components or subskills comprised the four aptitude areas.

Several different and varied specialties were then selected in order to test the approach and definitions. It became quite apparent early on that the four aptitude categories were too broad to adequately describe the components or work characteristics of a job or specialty. In most cases, it became necessary to break down the four broad categories into smaller, more meaningful categories.

For example, in the ADMINISTRATIVE area, it was discovered that there are at least three types of administrative tasks. First, there are those which involve clerical work, such as filing, preparing and maintaining forms and publications, and answering telephones. Second, there are tasks which deal with some form of mathematical computations, such as those performed in accounting or finance. And third, there are those tasks which involve the use of office equipment, such as typewriters, copy machines, or stenographs. Thus it became necessary to use three categories in this area rather than one.

In the MECHANICAL area, it was found that not all mechanical tasks were of equal weight. Some tasks were simple and only involved the use of such simple or common tools as a hammer or screwdriver. Other mechanical tasks were found to be somewhat more complex and involved spatial reasoning or advanced knowledge of a system in order to perform. Thus, some distinction was made as to relative difficulty of these tasks. Also, equipment operation (other than office equipment) was considered a mechanical skill. But then again, there had to be some distinction made between simple equipment operation, such as driving cars and vans, and more complex equipment operation, such as operating cranes, bulldozers, or aircraft K-loaders. Thus, further breakdowns were essential.

In the ELECTRONIC area, it became necessary to make some distinction between tasks that were purely electronic and those that were purely electrical, since there is a difference in the skills and knowledges required to perform tasks in either area. Also, in this area as well as in the MECHANICAL area, some distinction was made to differentiate those tasks that involved a combination of skills or knowledges, such as Mechanical-Electronic, Electrical-Mechanical, Electronic-Mechanical, and Electronic-Electrical.

The GENERAL category presented some problems in that it was more or less a "catch-all" category for those areas which were not involved with the other three categories. Most of the subcategories listed here included simple physical labor, medical skills, communicative skills (both oral and written), general procedures or techniques, planning, reasoning, and analyzing, scientific skills, and special talents (such as illustrating). But even here, it was necessary to provide some further delineations. For example, in the Medical subcategory, tasks were found to relate to either patient care or patient interaction, medical lab equipment operation, or medical procedures conducted in a medical lab or operating room. Thus, three additional subcategories were listed.

In all, 26 subcategories in four broad areas were finally listed. For each of the 26 components or subcategories, benchmark tasks were listed. (For a complete list of the categories and subcategories, interested personnel can write for a copy from the USAF Occupational Analysis Program (OMY), USAF Occupational Measurement Center, Randolph AFB, Texas 78150. Attention: Dr. Driskill).

DATA GATHERING

Subject matter specialists (SMSs) TDY to the USAF Occupational Measurement Center to write Specialty Knowledge Tests (SKTs) were used in categorizing tasks from the various specialties. For each specialty to be reviewed, those tasks which comprised 50 percent of the total job time for the journeyman (5-skill) level were selected. This was an arbitrary percentage which was felt to give a good representation of the technical tasks performed by the population in a given specialty. In addition, task difficulty data routinely collected on each specialty were used to help categorize those tasks where difficulty was a factor.

As each team of SMSs were used, they were explained the categories and shown the benchmark tasks for each. Definitions were carefully explained to them in detail so as to avoid any confusion. As each team went through the task lists, they were asked to explain what was involved in performing the tasks, what type of skills or knowledges were involved, etc. Occupational analysts used their comments in deciding the category which best fit the established benchmarks. In most cases, a single category was appropriate for any given task. In some cases, however, a task would involve multiple categories or components. For example, in the Small Arms specialty, many of the tasks are general in nature, since they involve instruction on the use and maintenance of weapons. But there is also a mechanical element to most tasks, since many of the steps shown in the instruction involved performing mechanical operations. When this happened, the task was placed in both categories.

Once all tasks were categorized, the totals for each category were tallied and a final overall category (Administrative, General, Mechanical, or Electronic) was listed for the specialty, along with other pertinent findings or components. The SMSs were asked their general opinion on the current AQE/ASVAB category and on the categorization results for their AFSC. Where there were differences found between the current AQE/ASVAB category and the USAFOMC categorization, they were asked their opinion as to which category they believed was most appropriate. This exercise tended to further validate the results of the project.

CONCLUSIONS

In looking at some 87 specialties to date, it is quite evident that a project of this nature is needed. While the four AQE/ASVAB categories currently in use are sufficient for many ladders, they are not sufficient for describing others. The nature of work performed today in some enlisted specialties has changed drastically from that performed years ago when the original AQE/ASVAB categories were established. As USAFOMC completes categorization in each of the approximately 200 major enlisted specialties, the information is being turned over to the Air Force Manpower and Personnel Center's classification Branch. Where differences exist between the current ASVAB classification and the USAFOMC categorization, a reexamination of the ASVAB testing category may be required.

By using a task-based approach to examine the types of work being performed in specialties, it is possible to provide better selection and recruitment information.

REFERENCES

Massey, U.H., Creager, J.H. Validation of the Airman Classification Battery: 1949-1953. AFPTRC-TN-56-29, November 1976

Mullins, C.J., Earles, J.A., Ree, M. Weighting of Aptitude Components Based on Differences in Technical School Difficulty. AFHRL-TR-81-19, August 1981

Vitola, B.M., Mullins, C.J., Croll, P.R. Validity of Armed Services Vocational Aptitude Battery, Form 1 to Predict Technical School Success, AFHRL-TR-73-7, July 1973

AD P001312

INTERACTIVE VIDEO FOR INTERPERSONAL SKILLS TRAINING[1]

Frederick N. Dyer and James E. Schroeder
Army Research Institute Field Unit, Fort Benning, Georgia

Paul Czerny, Daniel P. Gillotti and Edward W. Youngling
Litton Mellonics Systems Development Group
Fort Benning, Georgia

ABSTRACT

This new interpersonal skills training technology uses a videodisc player controlled by a microcomputer. The videodisc depicts a number of possible interactions between a new Army lieutenant and his subordinate which might occur when the lieutenant attempts to solve a problem such as deficient subordinate performance. The leadership trainee is first presented background information related to the problem. The trainee then sees and hears the subordinate's initial comment on the television monitor. A menu of possible responses the student might make follows on the TV screen and the trainee selects the one he feels is best by pointing to it with a light pen. The computer program causes the videodisc player to move to the point on the videodisc that depicts the way the subordinate might react if treated in that manner. The subordinate's reaction to a given response is designed to provide feedback about the quality of that response. One mode of instruction attempts to simulate an interpersonal interaction as closely as possible. Another adds additional feedback about the quality of each response, whether it is the best response, and the reason it is correct or incorrect. Initial reactions of individuals reviewing the first of eight videodiscs have been highly positive. An experimental evaluation of its training and assessment potential begins in November 1981.

INTRODUCTION

Background

Previous research at the Army Research Institute Field Unit at Fort Benning, Georgia, showed that a videodisc system could successfully train soldier technical skills even when only a fraction of the potential of the videodisc medium was used (Holmgren, Dyer, Hilligoss, & Heller, 1979). The current research and development effort at the field unit more fully exploits videodisc technology by providing simulations of leader-subordinate interactions for realistic training of interpersonal leadership skills. These videodisc scenarios will allow new Army leaders to practice interactions with simulated subordinates in situations which now are frequently mishandled in actual Army settings.

This videodisc interpersonal skills training and assessment (VISTA) project was initially conceived as a way to reduce the high personnel costs associated with use of assessment centers for assessing and developing leadership skills.

[1]This paper was also presented at the American Psychological Association Convention and the Society for Applied Learning Technology Convention.

The problem was one of simulating human beings in the many different ways that they might respond in a leadership interaction. An audiovisual medium was needed that would allow rapid accurate random access to a large number of motion sequences. This could not be accomplished satisfactorily prior to the advent of the videodisc.

Participants in the Project

Three Army agencies, a civilian contractor, and two Army television studios are working with the Army Research Institute in the development and evaluation of this new technology for interpersonal skills training. The Army agencies are the Training Developments Institute of the US Army Training and Doctrine Command at Fort Monroe, Virginia which is funding scenario development and evaluation; The Army Communicative Technology Office at Fort Eustis, Virginia which is providing equipment and videodisc mastering; and the US Army Infantry School at Fort Benning, Georgia which is providing leadership subject matter experts. The Litton Mellonics Systems Development Group at Fort Benning, Georgia is developing the leadership training scenarios, integrating the computer and videodisc hardware and developing the computer software. They will also carry out the experimental evaluation of the materials. Video production is being done primarily by the Fort Benning Educational Television Branch with some assistance from the Training and Audiovisual Support Center at Fort Gordon, Georgia.

COMPUTER-VIDEODISC TECHNOLOGY

Videodisc technology provides television displays which are much more flexible than those which come from videotape. Access from any one frame or sequence to any other frame or sequence is less than five seconds for the videodisc player we are using. Less distant segments on the disc can be reached and displayed in less than one second. In addition, during forward and reverse searches, the videodisc player is monitoring the frame number. As a result, the exact frame can be selected which corresponds to the beginning of a new motion sequence. Alternatively, the videodisc player can repeat that single frame over and over again for a static display. (This single frame feature would allow a slide-show of 54,000 separate pictures.)

The player can be interfaced to an external microcomputer. This allows computer controlled branching to different segments of the videodisc. It also makes it possible to present computer graphics on the video screen for increased instructional flexibility. In addition, a light pen or touch panel can be added to the system to permit the student to interact with the display. Finally, a real-time clock can be used to measure response latencies and use those latencies as cues for certain video segments.

Our system consists of an MCA PR-7820 videodisc player interfaced to an Apple-II computer (48K plus PASCAL language card) via a Colony Products VAI controller card. We also use a Symtec light pen and a Mountain Hardware real-time clock. All software is written in the programming language PASCAL.

LEADERSHIP TRAINING APPLICATION

Two different modes of instruction have been developed. The "experiential" mode will be discussed first followed by the "pedagogical" mode.

In the typical training situation, a new junior officer leadership trainee sits before a television receiver holding the "light pen" that allows direct interaction with the display. Typically, the television is first used to present some written background information to the trainee about a soldier who presents a leadership problem. This might be a new private in the platoon who has financial problems or an NCO who has been verbally abusing members of his squad. That individual then appears on the screen and is seen entering the lieutenant's office or approaching the lieutenant in the field setting. The televised subordinate typically begins the interaction, speaking directly to the viewer.

Following the background data and this initial comment by the simulated subordinate, the leadership trainee is shown a televised menu of possible responses that a new lieutenant might make in the situation. Each response was carefully chosen by scenario developers to appeal to at least some new leadership trainees. However, some of the responses are much more appropriate in the situation than others. The trainee reviews these alternative responses and points with the light pen to the one believed to be best. Immediately, the simulated subordinate reappears on the screen, behaving as he probably would if treated in the manner that was selected from the response menu. When this video segment depicting the simulated subordinate is complete, a new menu of responses for the trainee appears on the screen and the trainee selects a response for this updated situation with the light pen.

In this "experiential" mode of instruction, interactions continue between the leadership trainee and the simulated subordinate for as many as ten exchanges until the situation is resolved for better or worse. In the latter case, the simulated subordinate might be last seen on the TV bolting away from the lieutenant muttering about incompetent second lieutenants. Should the leadership trainee pause too long prior to responding, the computer would "know" this and would automatically display the simulated subordinate saying something like "If you are finished, Sir, I need to get back to the troops."

It is expected that these interactive scenarios with their rapid branching will cause trainees to react and respond to the subordinate depicted on the TV in much the same way as they would to a real subordinate. This approach to interpersonal skills training might provide a potent tool for training leader skills that unfortunately now are frequently learned only by trial and error on the job. In the instructional mode described above, the videodisc interactive scenarios will also provide trial-and-error learning, but the errors will not have serious negative consequences for the person the leader deals with or for the leader himself.

The second or "pedagogical" mode of instruction provides more feedback to the leadership trainee. In this mode, the student will first be asked to construct a response for a given leadership situation, by either writing it down or thinking it through. Next, the trainee is presented a response menu and asked to select the response that is closest to the trainee's answer, or is the best alternative. After selecting an alternative, the student has the option of previewing the response. If this option is selected, the trainee is shown the model lieutenant (actor) making that response. This preview insures that the trainee is not tricked by verbal-behavioral discrepancies. If the student does not like the response, the program branches back to the response menu. After the leadership trainee decides to keep a response, the videodisc plays the response (camera on subordinate's face), and the trainee sees the subordinate's reactions during and after the statement. Following this motion sequence, computer-generated text informs the trainee whether or not the response selected was the best option and provides precise feedback about why it was correct or not. If the alternative selected was not the best choice, the trainee is again presented the response menu but with the incorrect alternative removed. If it was the best alternative, the student is given the option of viewing any or all of the wrong alternatives to see why they were less appropriate. When finished with the first choice point, the trainee is taken to the second choice with a brief video review to recreate the situation. In this pedagogical mode the trainee is never allowed to go more than one step off the "best path."

The same videodisc can be used for both the "experiential" and the "pedagogical" modes of instruction, because the computer software dictates the mode of instruction. Research is planned to establish the optimal means for combining these two modes of instruction.

PROGRESS AND PLANS

The initial videodisc scenario was completed in the Summer of 1981. All eight will be completed less than a year later. The scenarios will receive their initial validation in a leadership course for new Infantry lieutenants. However, they might be sufficiently general that they could be used for other Army Branches and possible leaders in other services.

The eight scenarios might also provide a powerful and inexpensive leadership assessment tool. All eight videodiscs will provide nearly 100 opportunities to measure quality of leader responses. Such assessment data could possibly be used to aid in selecting candidates for the Army's Branch Immaterial Officer Candidate Course (formerly OCS) or for the Military Academy. The procedure might also be used as a voluntary refresher course for more experienced Army leaders.

Future videodisc developments are anticipated for training the critical interpersonal skills of race-relations officers, chaplains, military police, and senior officers. Training of tactics and combined training of tactics and interpersonal leadership skills are also foreseen.

Reference

Holmgren, J. E.; Dyer, F. N.; Hilligoss, R. L. and Heller, F. H. The effectiveness of Army Training Extension Course lessons on videodisc. J. Educational Technology Systems, Vol. 8(3), 1979-80.

ABOUT THE AUTHORS

Dr. Frederick N. Dyer is a Work Area Leader with the Army Research Institute for the Behavioral and Social Sciences. He is currently responsible for research on computer-assisted leadership assessment and training. He holds a Ph.D. degree from Michigan State University in Experimental Psychology.

Dr. James E. Schroeder is currently a Research Psychologist with the Army Research Institute for the Behavioral and Social Sciences. Prior to joining ARI, he was a Research Psychologist with Litton Mellonics and an Associate Professor at Lamar University where he conducted research in Learning and Motivation. He earned his Ph.D. in General Experimental Psychology at the University of New Mexico.

Dr. Paul Czerny is a Senior Operations Research Analyst with Litton Mellonics. He holds a Ph.D. in Biopsychology from the University of Georgia.

Mr. Daniel P. Gillotti is a Senior Research Analyst with Litton Mellonics. His former work assignment was as a Training Development Specialist with the Vought Aerospace Corporation. He is retired from the Army and holds an M.A. in Counseling Psychology from Ball State University.

Dr. Edward W. Youngling is Projects Manager for Litton Mellonics' Fort Benning operations. He holds a Ph.D. in Experimental Psychology from the University of Massachusetts.

TESTING RECALL BY MEANS OF MULTIPLE CHOICE TESTS

by

Ralph W. Ekwall, Ed. D.
Educational Specialist
US Army Command and General Staff College
Ft. Leavenworth, Kansas

This paper defines and discusses the difference between recall and recognition. Examples are presented to illustrate the difference. The implication of the difference is that modern training programs are computer managed, or computer assisted and make use of machine scored answer sheets. These modern systems have not been capable of testing recall-only recognition. The graduate of a training program needs to be able to recall rather than recognize. The system for testing recall by means of a machine scored answer sheet is explained and demonstrated.

TESTING RECALL BY MEANS OF MULTIPLE CHOICE TESTS

ABSTRACT

This presentation explains the difference between recall and recognition. The implications of the difference are discussed. Special attention is given to the fact that the extensive use of machine scoring causes recognition to be tested rather than recall. The method for testing recall by means of multiple choice testing is explained and an example is demonstrated.

EXPLANATION

Recall and recognition are different mental processes. Recall requires the stud to retrieve a term or word from memory. The student is provided no help; he is aske for a response and the student must go through the process of searching through his mental files trying to find the correct term or name. Recall is an active process and a higher level of mental activity than recognition. Recognition, on the other hand, tends to be a passive process. The student is required to select from a list presented to him. In summary: the processes are different; recall is a higher level activity with more requirement for mental activity; recognition is a lower level mental activity which tends to be passive.

At this point an attempt will be made to provide some evidence of the difference between recall and recognition. You may have had the experience of thinking that you know an answer to a question, but not being able to state the answer. When you see the answer among a set of responses you are able to identify the answer. This is an example of being able to recognize, but not recall. Let's try some examples. You will be asked a question; a few moments after the question is asked a transparen will be projected showing possible responses - one of which is correct.

Who was vice-president during the last four years of Lyndon Johnson's presidenti term?

A. Hubert Humphrey
B. Walter Mondale
C. William Miller
D. Spiro Agnew
E. Nelson Rockefeller

What is the scientific name for the dog?

A. Felis Catus
B. Equus Cabellus
C. Semper Fidelis
D. Canis Familiaris
E. Fidelis Fedelis

What is the name of the outermost planet of our solar system?

A. Saturn
B. Mars
C. Jupiter
D. Pluto
E. Mercury

Next, we will look at the implications of the difference between recall and recognition. The distinction is often fuzzy in the world of training. We often write objectives that say "the student will identify..." and this often means recognize because the student is required to recognize the correct term in a 4 choice multiple choice question . If the objective says "the student will recall...," then the student should be required to recall - not to recognize. In real world working situations the distinction is important. In the real world, workers are required to recall-not recognize. The ordnance technician who needs a new cam pin should be able to recall the name of the part; he will not be given a list from which he will recognize the needed part. He must recall and then say or write the name of the needed part; it is not enough for the worker to recognize names of parts. The essential point is this: It may be more desirable to train students to recall terms rather than training for recognition.

Modern training programs have been developed to achieve a maximum of efficiency. To achieve this degree of efficiency, training programs often rely on testing and evaluation systems that utilize machine scoring, computer assistance and computer management. This means that multiple choice items, which can be machine scored, are used extensively in modern training programs. Since extensive use is made of multiple choice items, most of the testing is testing of recognition rather than recall. There is a need to be able to test recall by means of multiple choice items. The remainder of this presentation will describe a method for testing recall by means of the multiple choice test.

The method requires that the student recall and write out the required response on an initial response sheet. This response will be the basis for answering a multiple choice question. The multiple choice questions will ask the student to identify certain numbered letters in the answer. An example is needed to illustrate the definition. Suppose that the question requires the word multimeter as the correct response. The student writes out the word multimeter on the initial response sheet. The question asks the student to identify the 2nd and 4th letters of the term and is given responses such as these:

A. n,t
B. c,n
C. r,s
D. u,t

The student counts the first four letters in the word multimeter and identifies u and t as the 2nd and 4th letters. The student therefore selects choice D as the correct answer. It may be helpful to go through another, more detailed example. Suppose that you are teaching about the parts of a master brake cylinder. Rather than dealing with every possible part, only six parts will be discussed. The parts are as follows. (See incl 1)

3. spacer
5. body
6. filler cap
10. piston cup
11. piston
15. push rod boot

Let's review each of the six parts to fix them in your mind...
Next, let's take a look at what some questions about these parts might look like. The first page we will call the initial response sheet. It requires the student to write in the names of the parts. Remember that the initial response sheet is not turned in for evaluation, but is used as the basis for answering multiple choice questions. (See incl. 2) The instructions preceding the multiple choice questions would be as follows. After you have written the names of the parts in the blanks, use that information to answer questions that follow. If the part name consists of 2 or more words, treat it as a single long word.

Question 1. Identify the 2nd and 5th letters of term 3.

A. r,n
*B. p,e
C. o,e
D. s,t

Question 2. Identify the 6th and 8th letters of term 6.

*A. r,a
B. s,e
C. o,r
D. n,l

This concludes the example. Note that in question 2 the term filler cap was considered as one word; the sixth letter was the last letter of the first word and a is the 8th letter of the 2 two words are combined.

DEMONSTRATION

Now let's try the system in something like a real world training situation. The transparency shows parts of an older model of flame thrower. (See incl 3) Notice that there is large spring called an adjusting spring enclosed by a spring case. At the top end of the spring is the adjusting spring botton. The spring pressure can be adjusted by the adjusting screw at the top of the mechanism. The opposite end of this pushes against the diaphragm assembly. There is an inlet and an outlet with the outlet apparently larger. There is also a smaller spring called the compensating spring; this spring is on the opposite end from the adjusting spring. Some other parts are the nozzle, the operating pin, and the body. A total of 11 parts are identified.

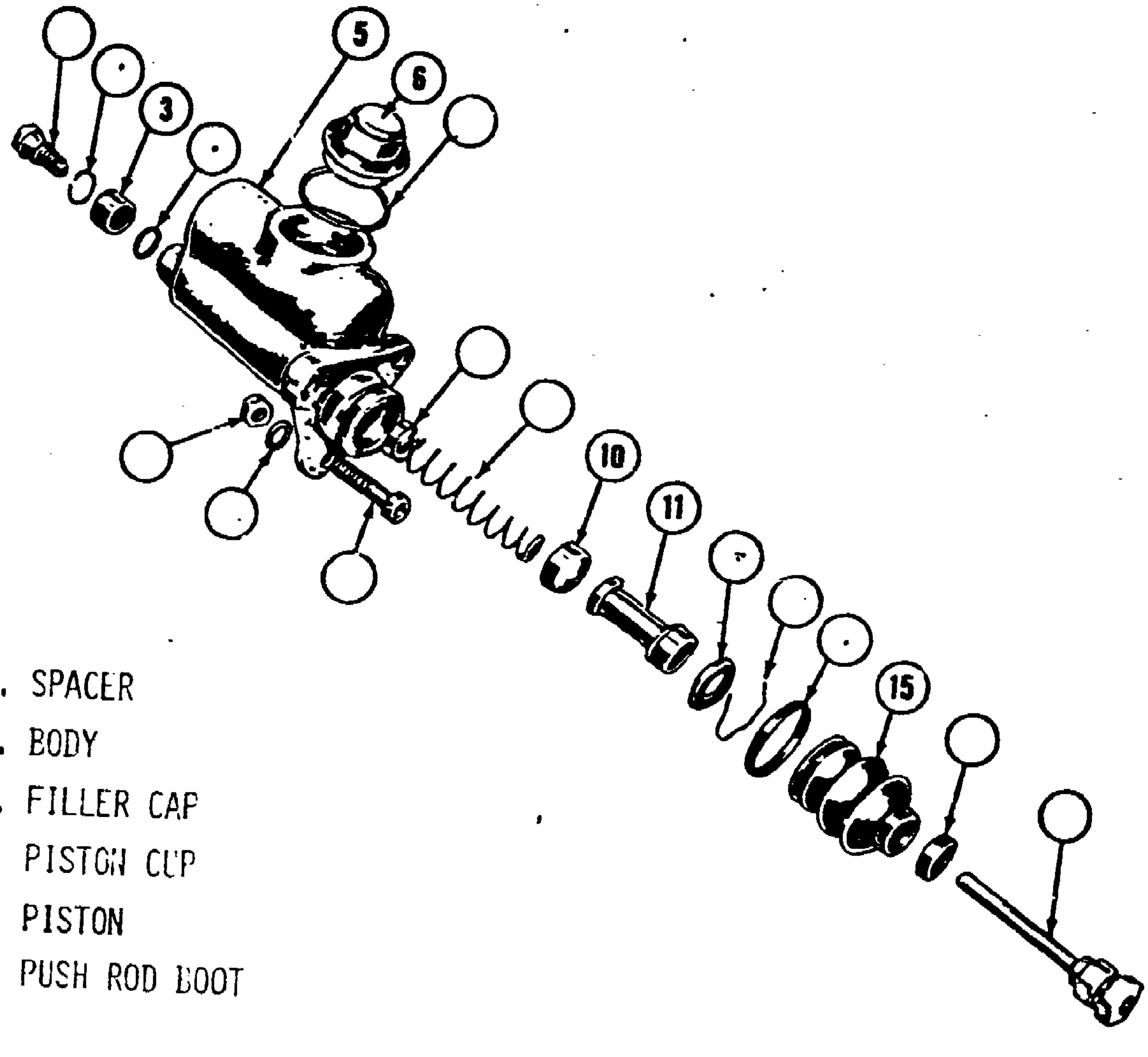
You will be provided with 2 documents. (See incl 4 & 5) The first is the initial response sheet which is a duplicate of the transparency. The other is the question sheet with a reproduction of a part of an IBM answer sheet. Fill out the initial response sheet and use this information to answer the questions on the response sheet.

The correct answers are: 1, E; 2, A; 3, A; 4, D; 5, C; 6, B; 7, D; 8 C; 9, A; 10, D; 11, E.

An obvious reaction to this method is that it requires the student to be able to spell and therefore it is partly a test of ability to spell. This problem deserves comment. It is often possible to minimize spelling problems. Diaphragm is a difficult word to spell; in the example you were required to know only the first and second letters so it was possible to get the correct answer without being able to spell the word. All that was necessary was to know the first 2 letters. The demonstration contained another problem that should be commented on. In the example there were 3 parts beginning with adjusting; to make sure the right term was being tested in question 7, the student was required to string 15 letters together correctly. In an actual teaching-learning situation we usually sample knowledge rather than testing 100% of items taught. Therefore, in an actual teaching-learning situation we could have avoided this problem by not testing all 3 of these parts.

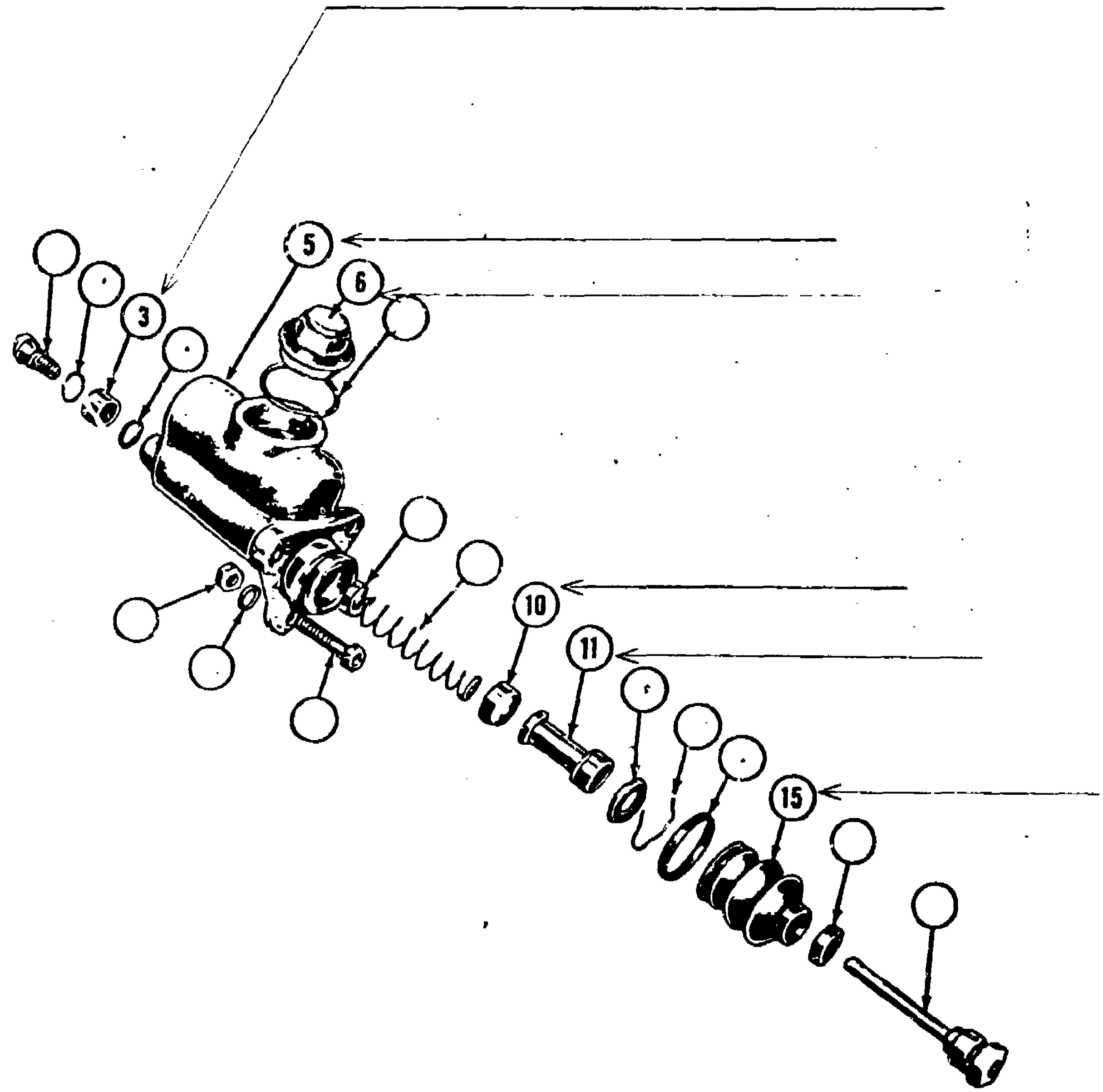
In the final section of this presentation some conventions are suggested. (1) Ignore capital letters; all responses will be small letters even though one of the letters of the answer is a capital letter. Example: the answer is Georg Simon Ohm and the 5 & 6 letters are required. Convention 1 says use g,s rather than G,s. (2) Treat multiple word responses as a long single word; do not count spaces between words. (3) Avoid requiring students to learn difficult spelling. If the spelling is difficult, ask questions about the first 2 or 3 letters of a word. (4) Avoid consecutive letters if feasible (5) Avoid use of first letters where feasible (6) Construct plausible alternatives; think of plausible alternatives and use letter combinations of those alternatives. If the correct answer is tinsnips and the numbers of the correct letters are (2,4) the correct response is (i,s). A plausible alternative is scissors and the plausible choice is (c,s). (7) Avoid asking questions about terms that have synonyms. This is not much of a problem in technical areas because technical terms seldom have exact synonyms.

A final note about this methodology: This is an experimental method; it has not been field tested, but informal experimentation shows promising results.

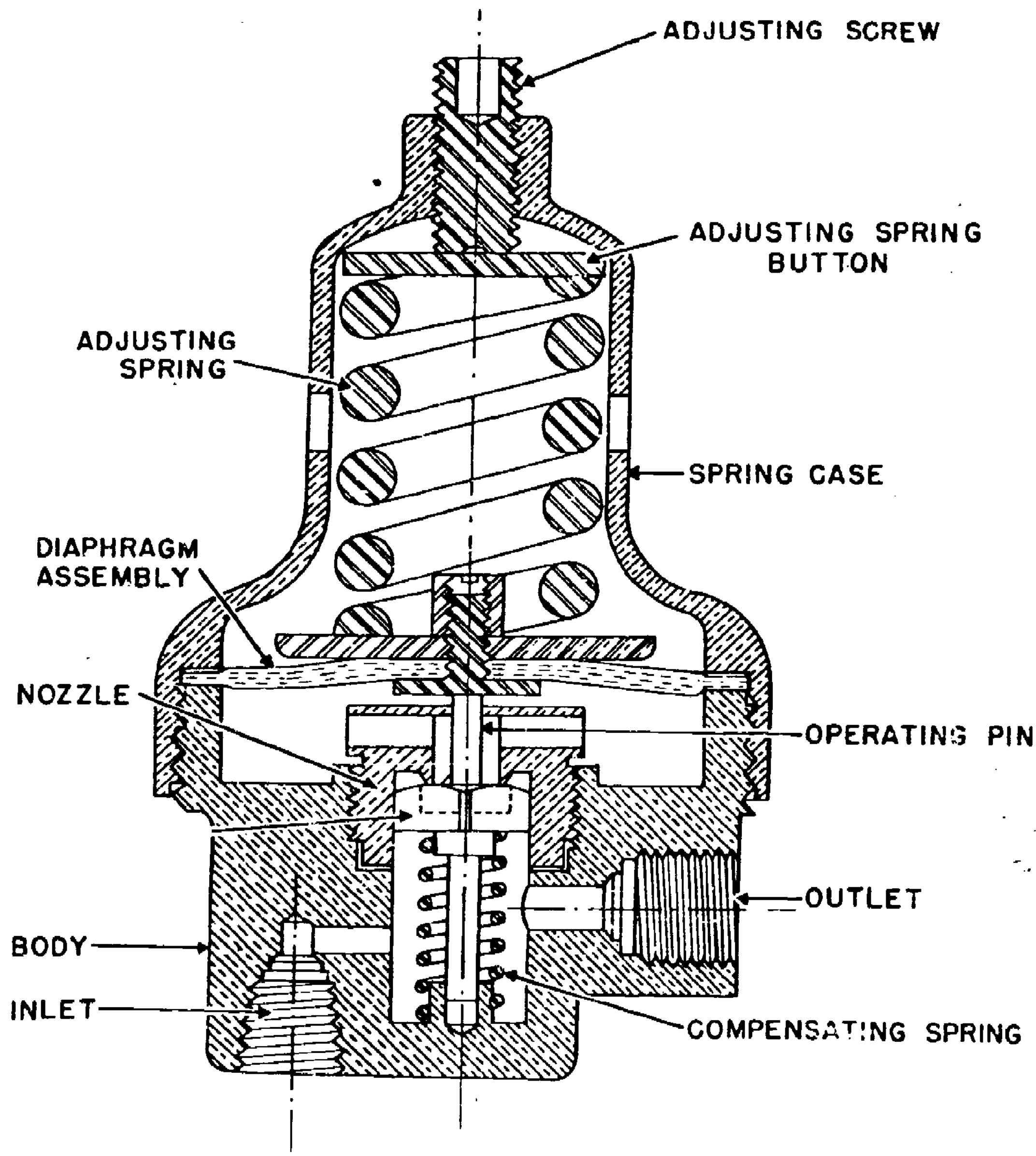


3. SPACER
5. BODY
6. FILLER CAP
10. PISTON CUP
11. PISTON
15. PUSH ROD BOOT

encl. 1

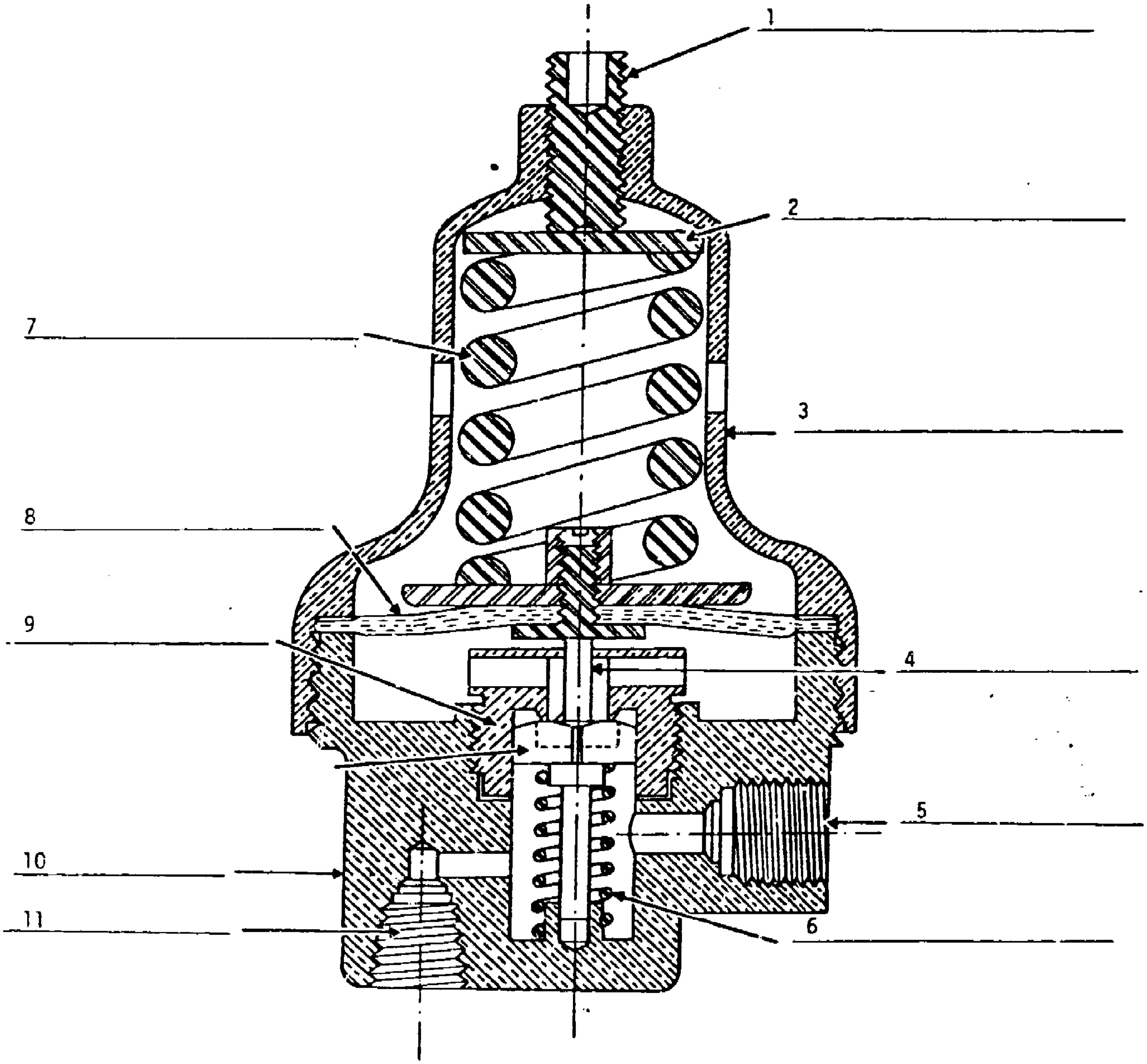


Encl. 2


ADJUSTING SCREW
ADJUSTING SPRING BUTTON
ADJUSTING SPRING
SPRING CASE
DIAPHRAGM ASSEMBLY
NOZZLE
OPERATING PIN
OUTLET
BODY
INLET
COMPENSATING SPRING

encl. 3

INITIAL RESPONSE SHEET



encl. 4

INSTRUCTIONS: Write out the identification of parts 1 through 11 on the initial response sh
Use this information to answer questions 1 through 11. Put your responses on
the IBM sheet below. If the part name consists of 2 or more words, treat it
a single word. The numbers in parenthesis after the question indicate the le
to be used in answering the multiple choice question. The 11 questions corre
to the 11 parts on the initial response sheet.

1. (2,14)
A. c,e
B. r,t
C. o,n
D. k,l
E. d,w

2. (5,11)
A. s,p
B. o,n
C. e,f
D. s,s
E. j,u

3. (2,5)
A. p,n
B. r,o
C. g,r
D. l,w
E. d,e

4. (3,5)
A. i,n
B. k,r
C. n,s
D. e,a
E. r,t

5. (2,6)
A. g,s
B. e,e
C. u,t
D. r,s
E. n,l

6. (4,6)
A. c,k
B. p,n
C. r,n
D. s,l
E. s,a

7. (3,15)
A. f,n
B. n,r
C. o,l
D. j,g
E. d,n

8. (1,
A. p,
B. b,
C. d,
D. c,
E. s,

9. (2,5)
A. o,l
B. n,r
C. s,w
D. c,r
E. o,e

10. (2,4)
A. l,r
B. s,r
C. n,r
D. o,y
E. d,e

11. (3,5)
A. e,a
B. j,l
C. r,t
D. o,f
E. l,t

T F	T F	T F	T F	T F	T F
1	31	61	91	121	151
2	32	62	92	122	152
3	33	63	93	123	153
4	34	64	94	124	154
5	35	65	95	125	155
6	36	66	96	126	156
7	37	67	97	127	157
8	38	68	98	128	158
9	39	69	99	129	159
10	40	70	100	130	160
11	41	71	101	131	161
12	42	72	102	132	162
13	43	73	103	133	163
14	44	74	104	134	164
15	45	75	105	135	165
16	46	76	106	136	166
17	47	77	107	137	167
18	48	78	108	138	168
19	49	79	109	139	169
20	50	80	110	140	170
21	51	81	111	141	171
22	52	82	112	142	172
23	53	83	113	143	173
24	54	84	114	144	174
25	55	85	115	145	175
26	56	86	116	146	176
27	57	87	117	147	177
28	58	88	118	148	178
29	59	89	119	149	179
30	60	90	120	150	180

encl 5

AD P001314

Elig, Timothy W., Gade, Paul A. & Eaton, Newell Kent, US Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia. (Wed. A.M.)

Performance Criteria Development for Army Field Recruiters

A variety of measures which have served as criteria of recruiter performance are discussed. New approaches to productivity measurement are developed to reflect both the relative value of different recruits to the Army and the influence of area fertility on recruiter productivity. Recent ARI research on FY79 productivity of 612 Army recruiters is presented. The large influence of District Recruiting Command fertility on individual recruiter productivity (accounting for 32% of the variance) was found to be primarily due to low priority recruits (those recruits who have low AFQT scores and/or did not get a high school diploma): DRC average production accounts for 34% of the variance in production of low priority recruits while it accounts for less than 9% of variance in production of high priority recruits. Managerial implications of area fertility adjustments of recruiter production are discussed. Recruiter reactions to performance appraisal adjustments for DRC fertility are considered.

Performance Criteria Development for Army
Field Recruiters

Timothy W. Elig, Paul A. Gade, and Newell K. Eaton

U.S. Army Research Institute for the Behavioral and Social Sciences
Alexandria, Virginia 22333

For both day-to-day operations and for long range planning, the development and utilization of recruiter performance criteria is of vital concern to managers of recruiting forces. Under the all Volunteer Force, the role of the military service recruiter has increased in importance and recruiting managers have felt the need to improve recruiter productivity. A key issue in improving recruiter productivity is how to measure recruiter productivity. We are concerned in this paper with two major changes in recruiting that have caused us to re-evaluate the way individual recruiter performance is measured.

The first major change is related to the increased importance of the individual recruiter since the cessation of the draft. Department of the Army demands placed upon the Army Recruiting Command are adjudicated thru three levels of this Command and ultimately placed on individual recruiters through monthly recruiting requirements. With no draft to make up shortfalls and to motivate individuals to enlist, the Command's concern with individual recruiter productivity increased exponentialy. Thus, performance criteria for recruiters narrowed from a broad concern with the recruiter as a soldier representing the military in a civilian community to a focus on the number of enlistments each month he or she could produce. Congressional concern with the quality of enlistees in the Army has been translated by the Recruiting Command to monthly recruiting requirements assigned to each Army recruiter. The monthly mission box assigned each recruiter is a three dimensional matrix of the number of Non Prior Service individuals the recruiter is to contract by recruit gender, education level and Armed Forces Qualification Test (AFQT) category, as well as a separate category of Prior Service applicants. Education level specifies individuals as being High School Diploma Graduates (HSDG), High School Seniors (HSSR), or Non High School Graduates (NHSG). AFQT category specifies whether the person is at or below the 31st percentile on the Armed Forces Qualification Test.

The increased emphasis on individual recruiter productivity is the concern of researchers both from the perspective of developing criteria for recruiter management research (e.g., efforts to find improved recruiter selection and assignment factors) and from the perspective of the understanding and acceptance individual recruiters have of the performance criteria used to evaluate them.

Role ambiguity and conflict which can result from the setting of performance standards are important concerns in personnel management. Role ambiguity refers to the degree to which an individual actually understands

---

The views expressed in this paper are those of the authors and do not necessarily reflect the view of the U.S. Army Research Institute or the Department of the Army.

what is required on the job. This is different from role conflict in which the individual understands the competing demands which are being made but may be unable to resolve which demands are more important. Role ambiguity and conflict have been found to be related to negative states such as dissatisfaction, stress, impaired performance and inappropriate organizational behavior (Rizzo, House, & Lertzman, 1970; Schuler, Aldag, & Brief, 1977; Keller, 1975).

The second major change that is leading to re-evaluation of performance criteria for recruiters is the need to develop criteria which are truly reflective of an individual's performance, and not merely reflective of large differences in task difficulty associated with geopolitical and socioeconomic factors of the recruiter's assigned area. For example, Bennett and Haber (Note 1) found that an urban or rural assignment was an important variable determining enlistment success of Marine recruiters. In another study of Marine recruiters, Larriva (Note 2) found that multiple correlations between various predictors of recruiter success and evaluations of recruiter's performance improved when geographic and rural vs. urban characteristics were controlled. Criterion research conducted for the Army during 1973 and 1974 was broadly focused on differences in recruiting "territory fertility". To account for fertility differences, individual performance was expressed as a deviation from the mean performance in the individual's territory (Fischl, Note 3). Average number of recruits per recruiter in a District Recruiting Command (DRC) was shown to account for 48% of the criterion variance of number of accessions (Brown, Wood & Harris, Note 4). While these findings demonstrate the importance of taking geopolitical and socioeconomic variables into account in criterion development, research remains to be done on exactly which predictor variables are important to measure and how to use them in criterion development.

Demographic variables are currently used in market analyses to determine the number of Qualified Military Applicants (QMA) in each recruiting station's area. The QMA is used to determine the mission requirements for each station. Thus area socioeconomics indirectly influence one performance criterion: the percent of mission objective the recruiter actually achieves. Here again we see a strong potential for role ambiguity and lack of acceptance of the relatively subtle way that "fertility" now influences the establishment of performance criteria. An even greater potential for conflict exists if performance criteria are explicitly based on area fertility.

In this paper we focus on how field recruiters and their immediate supervisors (recruiting station commanders) feel about the current performance criteria used for Army recruiters and how they feel about alternative criteria that might be used. Specifically, we investigated the level of recruiter and station commander understanding and acceptance of current criteria, their preference for other criteria and their reactions to adjusting performance on the basis of DRC "fertility."

Methods

Data reported in this paper are preliminary. They include only 22 of 50 recruiting stations to be contacted. This data collection effort will be completed in early November 1981. Complete details of subject selection and all procedures can be obtained from the authors.

Respondents were 44 recruiters and 22 station commanders from the Western, Midwestern, and Southwestern Army Recruiting Regions of the United States. Two recruiters and the station commander were individually interviewed in each of two stations in each of eleven DRCs.

Performance measurement was the first substantive issue covered in all these interviews. After the respondents were asked about problems encountered in filling out our questionnaires and suggested improvements in the surveys, the interviewer raised the topic of performance rating. Respondents were asked to read a description of a modified performance rating system that could be used to compare recruiting performance of recruiters in the different DRC's (see Appendix 1). After being given the DRC Correction[1] for their DRC each respondent was asked how they would react to such a system being used.

Respondents were then asked to describe how they believe field recruiter performance is rated now. This was followed by a question asking recruiters and station commander how they would like to see field recruiter performance rated. Finally, respondents were asked to pick one measure as the best measure of field recruiter performance. The measures they were asked to choose from had been included in their surveys. Station commanders had previously rated each recruiter by an experimental performance report which included the questions in Appendix 2. Each recruiter had also rated themselves on these items.

Results

Information gathered in the interviews of recruiters and station commanders are presented below. We first present the information gathered on respondent's reactions to a modified performance rating system which adjusted contract totals for DRC fertility. Next we present how station commanders and recruiters believe recruiters' performance is currently rated and how they would like to see it rated. Finally, we present the measures recruiters and station commanders feel are the best measures of recruiter performance.

<u>Reactions to a Modified Rating System</u>

Overall, 23% of respondents (18% of commanders, 26% of recruiters) accepted the system of DRC corrections. The largest group of respondents (62% overall, 77% of commanders and 53% of recruiters) rejected a system of DRC corrections because it did not go far enough and adjust for "within DRC" fertility differences. Thus 85% of respondents indicated an initial

---

[1] DRC corrections were computed on a 1979-80 base for a six month production period as follows: $\text{Correction}_i = 6(\overline{DRCAV} - DRCAV_i)$, where $DRCAV_i$ is the DRC's average monthly contract production per recruiter in a ten month base period.

acceptance of some type of area fertility adjustments. However, 15% of the recruiters rejected fertility adjustments on the basis of their assertion that productivity depends only upon the recruiter--his or her effort or sales ability. Nine percent of the respondents (14% of commanders and 7% of recruiters) indicated that the system was unfair because it would take contract credit away from recruiters in DRCs above average in productivity. This concern was not raised by the other respondents.

Current and Preferred Rating Systems

In the performance measurement interview, respondents were asked how field recruiter performance is now rated and how they would like to see it rated (the second and third interview questions respectively). Responses to these questions were coded in two ways. First, respondents' views of current and preferred rating methods were coded for agreement or disagreement. Second, the current and preferred rating methods were each coded in three specific ways. The specific codings for each were: a) the relative importance of production numbers versus other performance criteria; b) the relative importance of total contract production versus categories of enlistees in the mission box; and c) type of criteria other than production numbers (e.g., recruiter effort).

Of the 64 respondents who could be coded, 70% (74% of station commanders, 68% of recruiters) described a preferred rating of field recruiter performance which was different from the system which they believe is currently used.

Table 1

Perceived Importance of Criteria of Field Recruiter Performance by Percentage of Respondents

Criteria Importance Coding	Current		Preferred	
	Station Commanders n=22	Recruiters n=42	Station Commanders n=20	Recruiters n=40
Production numbers only	77	81	30	45
Numbers and other criteria - Numbers more important	14	14	0	8
Numbers and other criteria - Equal Importance	4	2	10	22
Numbers and other criteria - Other criteria more important	0	2	30	5
Other criteria only	4	0	30	20
	99%	99%	100%	100%

Table 1 shows that the vast majority of recruiters and station commanders believe that production numbers are the most important measure of recruiter performance under the current system. This table also shows that the respondents would prefer to de-emphasize production numbers in measuring recruiter performance. Furthermore, station commanders and recruiters differ in the emphasis they place on numbers in their preferred criteria ($\chi = 10.2$, 4df, $p < .04$). Surprisingly, recruiters are less likely to reject production numbers as a preferred criteria than are station commanders.

Table 2

Perceptions of Productions Numbers as Criteria for Field Recruiter Performance by Percentage of Respondents

Production Number Criteria Importance	Current		Preferred	
	Station Commanders n=22	Recruiters n=44	Station Commanders n=22	Recruiters n=44
Numbers not mentioned or said to be unimportant	13	2	48	30
Unspecified Production Numbers	39	35	35	30
Specified Production Numbers				
Total contracts only	9	28	9	16
Contracts and mission box - Contracts more important	13	9	0	12
Contracts and mission box - Equally important	13	14	4	5
Contracts and mission box - Mission box more important	0	5	0	0
Mission box only	13	7	4	7
	100%	100%	100%	100%

Table 2 presents the respondents view on production figures as performance criteria. The first row of this table reiterates the previous finding that the respondents preferred ratings less dependent on contract production figures than they perceive the current ratings to be The results presented in this table also indicate that while respondents agree on the importance of production figures as measures of performance, they do not agree on which particular production figures are important. This lack of agreement on production figures is found in both station commanders and recruiters in both their understanding of the current rating system and in their preferred rating system.

Table 3

Frequency of Criteria Other Than Production Figures

Criteria	Current	Preferred
Attitude	3	1
Appearance	3	0
Paperwork	2	0
Credibility, knowledge as recruiter	1	1
Effort, volume of applicants worked	4	10
Supervisor ratings	2	19
Quality of enlistees but not by mission box	0	5
Not on a month by month basis	0	4

The final coding of current and preferred ratings was for performance measures not using production figures as criteria. Table 3 lists eight criteria which were cited by two or more respondents. Four other responses were made only by one individual and are not listed here. Station commanders and recruiters differed only in the extent to which they preferred supervisory ratings. Only 14% of recruiters selected supervisory ratings as the preferred performance measure while 44% of the station commanders preferred this measure.

Best Measures of Recruiter Performance

The last question on performance measures asked respondents to look at the questions in Appendix 2 and choose one of these questions as the best measure of field recruiter performance. Table 4 presents the percentage of station commanders and recruiters who choose each suggested measure as the best measure of recruiter performance. Suggested measures can be grouped into six general categories as shown in Table 4. There were sharp differences between recruiters and station commanders on what recruiters can and should be held responsible for. Twenty-five percent of the respondents choosing an applicant processing measure emphasized that the recruiter lacks control over the quality of the people he/she processes and over whether an individual will contract. Thirty-six percent of the respondents chose "total contracts" or "contracts as a percentage of the contract objectives" as the best measure of recruiter performance because it is "what the job is all about". While many recruiters communicated a concern with the issue of "quality" recruits, only 22% of the respondents felt certain enough about a recruiter's responsibility and/or ability to influence recruit quality, to choose a quality indicator as the best measure of recruiter performance.

Table 4

Percentage of Respondents Choosing Each Measure as the Best Measure of Performance

Measures	Station Commanders n=22	Recruiters n=42
Overall Ratings		
1. 5-pt scale	5	2
17. 7-pt scale	0	0
18. Enlisted Evaluation Report	5	a
Applicant Processing		
2. Contacted for at least 20 min.	9	14
3. Test	5	7
4. Send for physical	0	2
5. Send for contracting	9	2
Contracts and Mission Objective		
6. Contracts	18	24
7. % of objective	9	17
Delayed Entry Program (DEP)		
8/9. Number of DEP losses	0	2
Objective Quality of Enlistees		
10. High School Diploma Graduates	9	0
11. AFQT I thru IIIa	0	0
12. HSDG and AFQT I thru IIIa	0	2
Subjective Quality of Enlistees		
13. Quality service for term of enlistment	5	10
14. They are right for Army	0	14
15. Army right for them	0	0
16. Become quality NCOs	0	0
	74%	96%
More than one response	27	2
	101%	98%

Note: Wording of measures for commander is in Appendix 2.

[a]This question was not on the recruiter self report

Discussion

Before examining the views of station commanders and field recruiters on new recruiter performance measures, it is best to examine their understanding and degree of acceptance of current performance criteria.

Role Ambiguity: Understanding of Current Ratings

Role ambiguity is concerned with the degree to which an individual understands what is required on the job. Recruiters and station commanders seem to be in agreement that field recruiter performance is currently rated primarily on monthly contract production. However, role ambiguity does seem to exist with respect to what extent current performance criteria are based on total contracts only or on achieving the mission box objectives. We cannot currently tell whether this ambiguity is due to individual recruiter or station commander lack of awareness or interest or to some recent shift in Command emphasis that has not been clearly communicated to the recruiters and station commanders.

Person-Role Conflict: Acceptance of Ratings

While many respondents stated that they did not know how field recruiter performance could be rated differently from the way it is currently rated, 70% of the respondents did express a preference for a rating method which differed from their perception of the current method. Station commanders in particular said they would like to see less use of production figures and more use of the criteria used in other Army assignments (e.g. supervisor ratings of the total person or soldier). Recruiters sought more recognition of their efforts and skills in working with applicants. These results indicate a role conflict among recruiters between their identity as U.S. Army soldiers and the performance criteria placed on them as U.S. Army recruiters.

A different sort of person-role conflict was expressed by the 17% of the respondents who choose a subjective enlistee quality measure as the best measure of recruiter performance. Many recruiters express a conflict between having to make total contract production numbers and their personal desire to enlist only individuals whom they feel are right for the Army.

Importance of Role Ambiguity and Conflict

The amount of role ambiguity and role conflict which we found warrants further efforts to understand the organizational environment of the field recruiter.

Indications of the importance of role conflict and role ambiguity in decreased organizational effectiveness are to be found in the general literature. Keller (1975) found employee dissatisfaction to increase with role ambiguity. Rizzo, House, & Lirtzman (1970) suggested that role conflict and ambiguity resulted in stress, and that this stress, in turn, resulted in dissatisfaction, poor performance and generally inappropriate organizational behavior. Schuler, Aldag, & Brief (1977) also found role conflict and ambiguity to be related to negative affective states such as dissatisfaction and stress.

All the recruiter's problems cannot be solved by eliminating role ambiguity or conflict. For example, establishing clear performance criteria based only on the single criterion of production numbers could push the recruiter toward malpractice. Such pressure might be reduced by placing more emphasis on performance criteria associated with being an honest, hard-working soldier.

Shifting Role Emphasis for Recruiters and Acceptance of New Measures

Much of the ambiguity in respondents perceptions of current performance ratings may be related to a newly emerging role for recruiters. Recent command emphasis on the quality of enlistees and recruiter "ownership" of enlistees seems to mark the beginning of a shift from the role of the Army recruiter as a seller to anyone who is willing to buy, to a role as a personnel recruiter seeking out the best applicants for the jobs the Army needs to fill. This new emphasis is seen by the respondents to conflict with the long standing criteria of total contract numbers. Recruiters report many conflicting demands being placed on them concerning the number and quality of applicants they should be seeking.

Development of Performance Criteria: Research Needs

Researchers in this area need to be aware of the current flux in recruiting performance criteria if they are to know the limits on their work. A recruiter who may be extremely successful in accomplishing the performance criteria as he or she perceives them may not be considered successful in accomplishing the criteria perceived by the researcher. Because of the emergence of recruit quality as an essential component of recruiter performance evaluation, research is needed on the effect of geopolitical and socioeconomic variables not only on total productivity but also on categories of productivity.

In the current research, we found that fertility adjustments are acceptable to most recruiters and station commanders if the adjustments are sufficiently explained and are done at a small enough level. These adjustments, however, would almost certainly be better done as adjustments to the standards of mission box objectives, rather than as adjustments to contract performance outcomes. While most recruiters and station commanders can accept the logic for area fertility adjustments in judging performance, they were less than enthusiastic about subtracting a DRC correction number from the number of contracts produced. Any system which reduces the number of contracts a recruiter is credited with, would be perceived as unfair and would probably lead to decreased motivation.

Reference Notes

1. Bennett, J.T., and Haber, S.E. Selection, deployment, and evaluation of Marine Corps recruiters. Technical Report for Project NR347020, Office of Naval Research). Washington, D.C. The George Washington University, School of Engineering and Applied Sciences, Institute of Management Science and Engineering, 1973.

2. Larriva, R.F. U.S. Marine Corps recruiter performance prediction study. Unpublished manuscript, 1975.

3. Fischl, M.A. Recent recruiter selection research in the U.S. Army. Paper presented at the American Psychological Association convention, Washington, D.C., 1976.

4. Brown G.H., Wood, M.D., and Harris, J.D. Army Recruiters: Criterion development and preliminary validation of selection procedure. Technical Report FR-ED-75-8). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, April 1975.

References

Keller, R.T. Role conflict and ambiguity correlates with job satisfaction and values. Personnel Psychology, 1975, 28, 57-64.

Rizzo, J., House, R.E., & Lirtzman, J. Role conflict and ambiguity in complex organizations. Administrative Science Quarterly, 1970, 15, 150-163.

Schuler, R.S., Aldag, R.J., & Brief, A.P. Role conflict and ambiguity: A scale analysis. Organizational Behavior and Human Performance. 1977, 20, 111-128.

EXPERIMENTAL PERFORMANCE RATING

This is only an Experimental Form. Our purpose is to obtain your comments on designing a system which is fair to all recruiters.

Many of the possible causes of recruiting performance are beyond the control of recruiters and are simply a part of the area an individual is recruiting in. USAREC is interested in the effects of expressing a recruiter's productivity in a way that is adjusted by the fertility of an individual recruiter's DRC. It is hoped that an adjustment to productivity based on DRC's fertility will allow a more accurate comparison of recruiters from around the country. If a recruiter's productivity is adjusted for his or her DRC average productivity, all recruiters in the country can be more easily evaluated on individual achievement.

One possible adjustment is to add to a recruiter's number of contracts the difference between the national average number of contracts and the average number of contracts in his/her DRC.

COMPUTING ADJUSTED PERFORMANCE SCORES

An Adjusted Performance Score (AP Score) expresses what your performance would be expected to be if you were recruiting in an area of average productivity. Adding a correction means that your performance would go up in an area of average fertility, since you are now working in a DRC of below average fertility. Subtracting a correction means that your performance would go down in a DRC of average fertility, since you are now working in an area of above average fertility and recruiting is easier than it is in an average DRC.

EXAMPLES

John is a recruiter in a productive DRC. Of the individuals he recruited and sent to an AFEES in the last six months, 21 signed enlistment contracts. His Adjusted Performance Score would look like this:

Number of contracts:	21
DRC Correction:	-5
Adjusted Performance Score	16

Joe is a recruiter in a DRC that is less productive than the average. He recruited 11 people in the last six months who signed contracts. Joe's AP Score would look like this:

Number of contracts:	11
DRC Correction	+6
Adjusted Performance Score:	17

If we do not take DRC fertility into account and just compare John and Joe on contract totals, John looks like a much better recruiter than Joe. But by using DRC Corrections to get Adjusted Performance Scores we can see that Joe and John are both producing at about the same level. In an area of average fertility we would expect very little difference in performance between Joe and John; the big difference we see between their current performance levels is attributed to how easy it is to recruit in John's DRC and how difficult it is to recruit in Joe's DRC.

Performance Report for Last Six Months

In this section there are a number of statements and questions about this recruiter's performance during the last six months. Think over the last six months and answer each question with your best estimate.

1. All in all, her/his job performance in recruiting is

   1. Excellent
   2. Above Average
   3. Average
   4. Below Average
   5. Poor

2. _____ In the last six months, how many prospects has he/she had at least twenty minutes contact with?

3. _____ In the last six months, how many applicants did he/she send to take the ASVAB?

4. _____ In the last six months, how many applicants did he/she send to an AFEES for a physical?

5. _____ In the last six months, how many applicants did he/she send to an AFEES to see a guidance counselor?

6. _____ In the last six months, how many people he/she recruited and sent to an AFEES signed a contract?

7. _____ In the last six months, what percentage of objective did he/she make?

OF THE SIGNED RECRUITS FROM THE LAST SIX MONTHS,

8. _____ How many dropped out of DEP?

9. _____ How many more do you think will drop out of DEP?

10. _____ How many counted for him/her as high school diploma graduates?

11. _____ How many scored above the 49th percentile (categories I thru IIIA) on the AFQT?

12. _____ How many BOTH scored above the 49th percentile AND were counted for her/him as a diploma graduate?

13. _____ How many do you expect to contribute quality service to the Army and not attrite before the end of their contract?

14. _____ How many do you feel are right for the Army?

15. _____ How many do you feel the Army is the right place for during the period of their contract?

16. _____ How many do you think are going to become the quality career NCO's the Army needs?

17. During the last six months, how would you rate his/her performance as an Army Recruiter?

    COMPLETE FAILURE 1 2 3 4 5 6 7 COMPLETE SUCCESS

18. _____ EER you would give him/her for last six months.

THE EFFECTS OF JOB SATISFACTION ON AIR FORCE ENLISTEE RETENTION

Kenn Finstuen, Ph.D. and Charles N. Weaver, Ph.D.
Manpower and Personnel Division
Air Force Human Resources Laboratory (AFSC)
Brooks AFB, Texas 78235

Abstract

The purpose of this research was to assess the concurrent and predictive validity of occupational attitudes, as measured by the Air Force Occupational Attitude Inventory (OAI), in relation to global job satisfaction, reenlistment intent, and actual reenlistment behavior of first-term enlisted airmen. The OAI was administered to two samples of airmen consisting of 1,217 personnel in 1973 and 4,784 personnel in 1975. Multiple linear regression equations were developed for each of the two years for samples based on reenlistment eligibility (from which two groups were formed: eligible only and eligible-ineligible combined) and on separation classifications (from which two more groups were formed: voluntary only and voluntary-involuntary combined). When considered as a set, the OAI items were found to bear a strong relationship to global job satisfaction, a somewhat lesser relationship to reenlistment intent, and a moderate, but highly consistent relationship to reenlistment behavior. These relationships maintained their significance when the baseline effects of 53 biographical and job-related variables were held constant in regression analyses. Findings indicated that global job satisfaction was associated with the following: job challenge, the use of airman abilities, and feelings of accomplishment. Reenlistment intent and actual reenlistment were most highly associated with satisfaction with pay and benefits as compared to civilian jobs, the consideration the Air Force gave enlistees, removal of irritants, and contributions to the national defense. Cross-validation of the 1973 and 1975 equations revealed that these relationships were stable through time.

I. THE AIR FORCE JOB SATISFACTION RESEARCH PROJECT

Since 1971, a comprehensive program of job satisfaction research has been conducted by the Manpower and Personnel Division of the Air Force Human Resources Laboratory. The objective of the program was to investigate the impact of work-related factors on job satisfaction and career decisions as a step toward reaching the goal of full utilization and retention of qualified personnel. The basic elements of this program were to (a) define and measure the dimensions of job satisfaction, (b) identify problem areas which had the greatest potential for improvement through job satisfaction research, and (c) assess the effects of job changes on job satisfaction attitudes and reenlistment decisions (Gould, 1976, p.5).

The first phase of the job satisfaction research project required that an inventory be developed to assess the dimensions of job satisfaction in the work environment of the Air Force (Tuttle & Hazel, 1974). In developing the inventory, Tuttle, Gould, and Hazel (1975) hypothesized relevant job satisfaction dimensions and produced a scale for measuring those dimensions. Gould (1978) validated the hypothesized dimensions, examined the rating scale, and reduced the item pool to a minimum number required to assess the job attitude domain of the Air Force work environment. The resulting inventory, the United States Air Force Occupational Attitude Inventory (OAI), is composed of three sections. Section I, General Information, consists of 51 items concerning Air Force members' biographical background and job-related information, and attitudes toward reenlistment, global job satisfaction, and job interest. Section II, Occupational Attitude Information, consists of 200 job satisfaction items, the last 10 of which apply to supervisory work and are completed only by airmen who supervise others as part of their job. The satisfaction attitudes of respondents are measured on a 9-point rating scale ranging from 1=extremely dissatisfied to 9=extremely satisfied. Section III, Importance of Job Aspects to Career Decisions, contains 35 items representing the dimensions initially hypothesized in the development of the inventory. The factors are also rated on a 9-point scale, ranging from 1=not important to 9=extremely important.

The OAI has been under development and refinement at various periods for over 8 years, and represents one of the most comprehensive and carefully researched job satisfaction measures of those commonly in use (Pritchard & Shaw, 1978). Since its development, the OAI has been used in a number of job satisfaction studies. Gould (1976) reviewed OAI-related research through September of 1976 and, since then, OAI-related research has included examinations of first-term and careerist attitude differences (Edwards, 1978) and differences among work roles (Finstuen, 1981).

II. PURPOSE AND HYPOTHESES

The purpose of the present study was to provide knowledge of the concurrent validity of the OAI against global job satisfaction and reenlistment intent, and to assess the predictive validity of the OAI against actual reenlistment. Four hypotheses were proposed:

- H1: Global job satisfaction, reenlistment intent, and actual reenlistment rates for first-term airmen will vary as a function of biographical attributes, job-related information, and occupational affect as measured by the OAI.
- H2: Functional relationships between the OAI and the attitudinal and behavioral criteria will be found to exist even when the effects due to biographical and job-related differences are controlled for or held constant in prediction.
- H3: The attitudinal saliency of specific OAI items displaying the highest degree of association with global job satisfaction, reenlistment intent, and actual reenlistment will remain stable across time.
- H4: Cross-validation of occupational attitude equations developed for samples in separate years will result in consistent and significant predictions of attitudes and reenlistment behavior across time.

III. METHOD

Subjects

An opportunity to examine the concurrent and predictive validity of the OAI was made available when the instrument was administered in 1973 and 1975 to random samples of enlisted Air Force personnel. Excluding career airmen, the samples included 1,217 and 4,784 first-term Air Force enlistees respectively for 1973 and 1975, for whom complete data were available.

Criterion Data Sets

Global job satisfaction, reenlistment intent, and actual reenlistment behavior were used as criteria. Global job satisfaction was assessed with the question, "In general, how satisfied are you with your present job?" Responses were made on an 8-point rating scale ranging from 1=extremely dissatisfied to 8=extremely satisfied. Reenlistment intent was measured by responses to the question "Do you plan to reenlist at the end of your current enlistment?" assessed on a 4-point rating scale ranging from 1=definitely will not reenlist to 4=definitely will reenlist. And, of course, reenlistment behavior was determined on the basis of whether an airman actually reenlisted or not.

While concurrent measures for both global job satisfaction and reenlistment intent were included in the 1973 and 1975 OAI surveys, it was not possible to complete the validation until the reenlistment behavior criteria had matured. In other words, sufficient time had to pass so that a reenlistment decision point was reached by all airmen in the samples. Furthermore, analyses were performed separately for airmen who were eligible to reenlist. This required that airmen who entered the service in 1975 be tracked for 36 months to the point at which qualitative screening for reenlistment eligibility could take place.

Reenlistment is one of three types of personnel actions which occur in the course of an airman's tour of duty. The other two types of actions are losses and extensions. Each event of these three broad categories is assigned one of over 200 three-digit Special Program Designator (SPD) identification codes. For the purpose of creating meaningful criteria, the SPD codes were classified on the basis of the type of discharge into voluntary and involuntary categories. A voluntary loss was defined as a separation initiated by the Air Force member. Examples of reasons for voluntary separations were to attend an educational facility, to accept public office, and to join a civilian police force. An involuntary loss was defined as a separation initiated by the Air Force. Examples of reasons for involuntary separations are shirking, sexual perversion, misconduct, and permanent physical disability.

Beyond division on the basis of voluntary-involuntary separation, the criteria were further classified on the basis of formal reenlistment eligibility. Eight attitudinal and reenlistment data sets were developed. For global job satisfaction and reenlistment intent, there were two categories, each based on eligibility. These are shown as criteria 1-4 below. For actual reenlistment, there were two categories of voluntary-involuntary separations for each of the two classifications of reenlistment eligibility. These criteria, 5-8 below, were dichotomously coded 1 if airmen reenlisted, and zero otherwise. The criterion data sets were:

1. Global job satisfaction: Eligible - Ineligible
2. Global job satisfaction: Eligible Only
3. Reenlistment Intent: Eligible - Ineligible
4. Reenlistment Intent: Eligible Only
5. Actual Reenlistment: Eligible - Ineligible, Voluntary and Involuntary
6. Actual Reenlistment: Eligible - Ineligible, Voluntary Only
7. Actual Reenlistment: Eligible Only, Voluntary - Involuntary
8. Actual Reenlistment: Eligible Only, Voluntary Only

Figure 1 displays the combinations of outcomes, discharge types, and eligibility classifications used for defining criterion data sets 5-8.

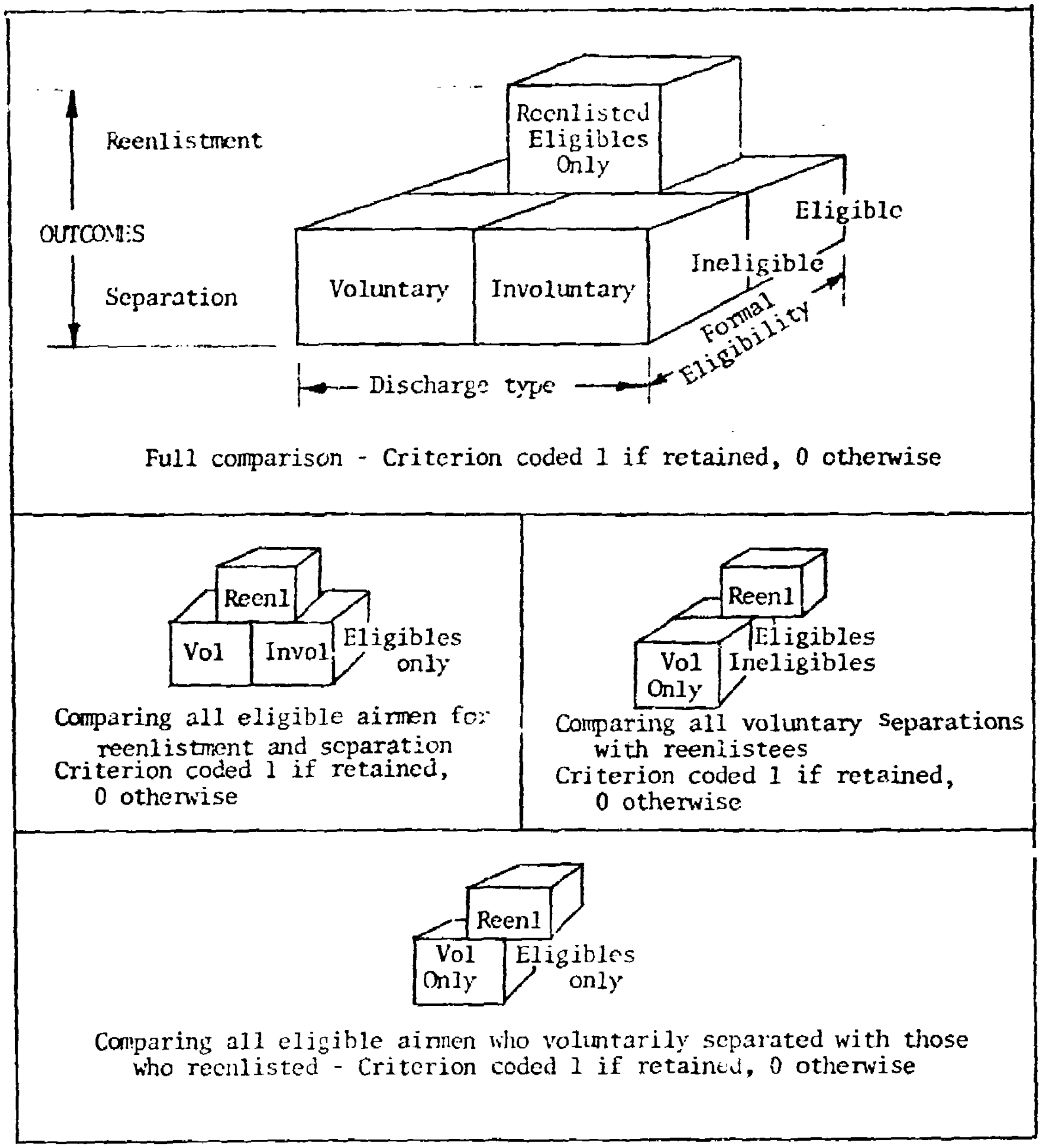


Fig. 1. Three-dimensional data structures for making predictions of reenlistment using various separation classifications.

The four criterion data sets portray various separation and reenlistment outcomes. Separations are classified by formal eligibility, either eligible to reenlist or ineligible, and by discharge type, either voluntary or involuntary. Those airmen that reenlist must be eligible. Obviously the type of discharge dimension does not apply to reenlistees. Both the 1973 and 1975 samples were coded as shown above.

Predictors and Equations

Two sets of predictors were included in the analyses: 189 non-supervisory items from Section II of the OAI and 53 pre- and post-enlistment baseline measures, including 25 biographical and 28 job-related control variables that are typically used in recruit selection, classification, and assignment actions. The biographical variables were Airman Qualifying Examination (AQE) aptitude scores, race, sex, age, education, marital status, number of dependents, size of hometown, and the amount of time spent reading. Job-related variables were months of total active federal military service (TAFMS), months on the job, number of subordinates, grade, and 18 occupational membership categories. Squared terms for aptitude, age, dependents, and number of people supervised were generated to detect curvilinear relationships when conducting the regression analyses.

Figure 2 portrays the functional relationships between the dependent and independent variables. Three kinds of multiple linear regression equations were developed (Bottenberg & Ward, 1963; Ward & Jennings, 1973). The first, or full model, equations were composed of pre- and post-enlistment control variables and the 189 OAI items. The second, or restricted model, equations were limited to the biographical and job-related variables. The third set of equations , also restricted, were based upon the OAI items alone.

RESULTS

Table 1 presents descriptive statistics for the 1973 and 1975 samples based on the various criterion data set classifications. Both attitudes and reenlistment rates were somewhat higher for the eligible only classifications. With the exception of the eligible only-voluntary-involuntary data sets, attitudes and reenlistment rates appeared similar for both 1973 and 1975. Overall, some 60 to 70% of the enlisted airmen did not reenlist.

Table 1
Criteria Development - Means and Standard Deviations for
Global Job Satisfaction, Reenlistment Intent, and Actual Reenlistment Rates

Criteria	Year	Eligible Only			Eligible - Ineligible		
		N	Mean	S.D.	N	Mean	S.D.
				Attitudes			
Global Job	1973	961	4.75	2.10	1,217	4.65	2.14
Satisfaction	1975	3,753	4.82	2.11	4,784	4.69	2.15
Reenlistment	1973	961	1.95	.84	1,217	1.91	.84
Intent	1975	3,753	2.29	.99	4,784	2.22	.99
			Reenlistment Behavior (Percent Retained)				
		N	%		N	%	
Voluntary -	1973	896	33.19		1,131	29.00	
Involuntary	1975	2,993	40.83		4,017	30.92	
Voluntary Only	1973	835	38.20		968	33.88	
	1975	2,988	40.90		3,650	34.03	

Note. Global job satisfacton was scaled 1=extremely dissatisfied to 8=extremely satisfied. Reenlistment intent was scaled 1=definitely will not reenlist to 4=definitely will reenlist.

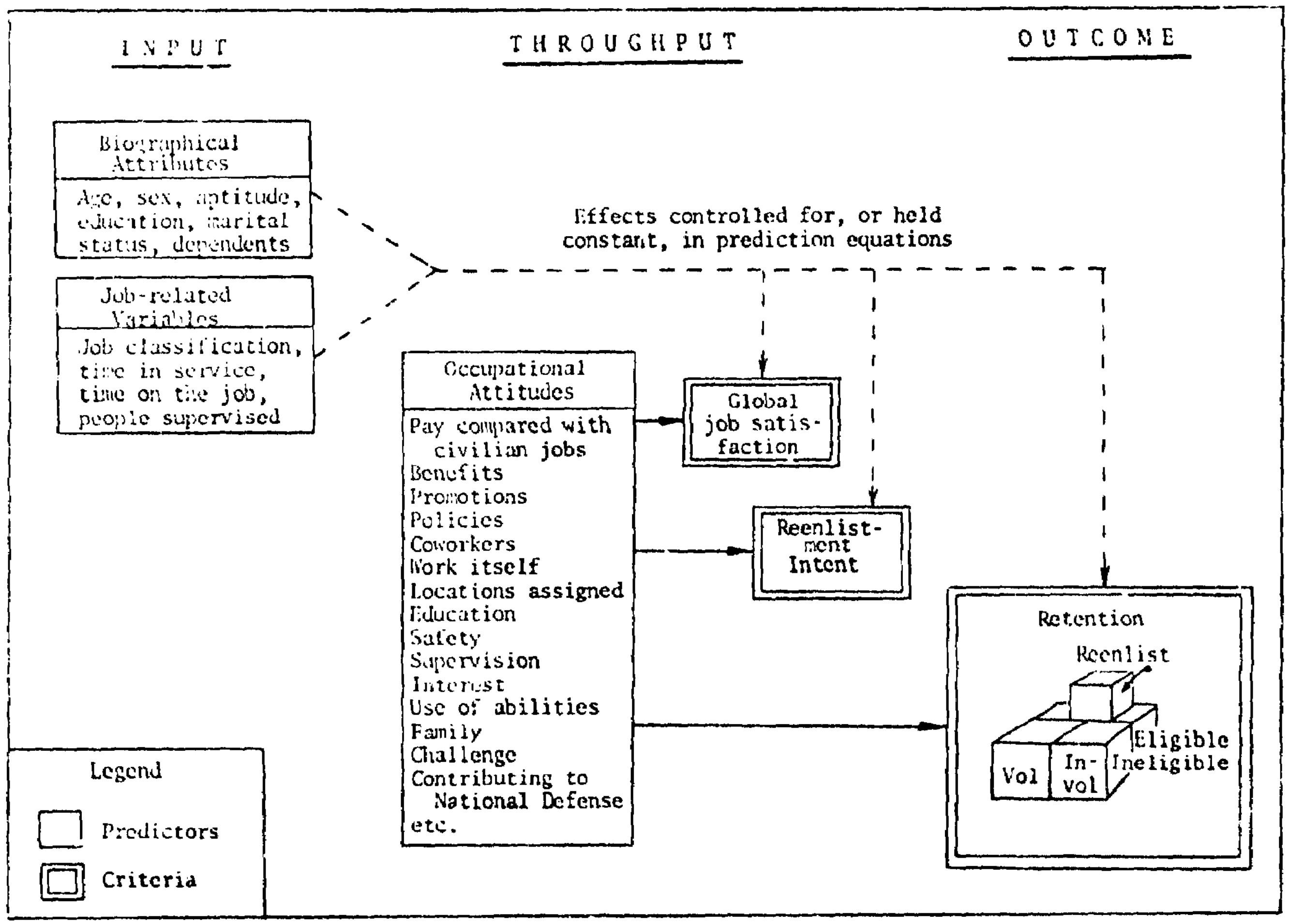


Fig. 2. Schematic diagram of dependent and independent variables used in the study.

Table 2 presents the multiple correlation results for all criterion data sets. In support of hypothesis 1, the statistically significant results for the full models (A) indicated that all criteria varied as a function of biographical, job-related, and occupational attitude variables. To determine the effects of the OAI item set, the multiple squared correlation coefficients

Table 2
Regression Analysis and Validation Summary
for Global Job Satisfaction, Reenlistment Intent, and Retention

Criterion	Full Models (A) R^2	Restricted Models (B) [OAI Removed] R^2	(A) vs (B) df_1	df_2	F^a	Restricted Models (C) [OAI Only] R^2
		1973 Survey				
Global Job Satisfaction						
Eligible-Ineligible	.71	.13	189	979	10.02*	.67
Eligible Only	.76	.17	189	723	9.38*	.74
Reenlistment Intent						
Eligible-Ineligible	.46	.16	189	979	2.86*	.41
Eligible Only	.51	.19	189	723	2.53*	.45
Retention						
Eligible-Ineligible						
Voluntary-Involuntary	.34	.15	189	893	1.36*	.24
Voluntary Only	.37	.14	189	730	1.37*	.28
Eligible Only						
Voluntary-Involuntary	.39	.16	189	723	1.46*	.30
Voluntary Only	.44	.17	189	597	1.56*	.35
		1975 Survey				
Global Job Satisfaction						
Eligible-Ineligible	.60	.09	189	4,546	31.21*	.59
Eligible Only	.61	.08	189	3,515	25.02*	.59
Reenlistment Intent						
Eligible-Ineligible	.33	.12	189	4,546	7.78*	.28
Eligible Only	.35	.12	189	3,515	6.55*	.30
Retention						
Eligible/Ineligible						
Voluntary-Involuntary	.21	.12	189	3,779	2.33*	.13
Voluntary Only	.20	.10	189	3,412	2.28*	.13
Eligible Only						
Voluntary-Involuntary	.21	.10	189	2,755	2.15*	.16
Voluntary Only	.22	.10	189	2,750	2.15*	.16

[a]All F tests comparing full (A) and restricted (B) models were significant* $p < .01$.

Note. Full models (A) contain OAI, biographical, and job-related variables. For restricted models (B) the OAI items have been removed and restricted models (C) contain only OAI attitude scores. All full model (A) and OAI model (C) coefficients were statistically different from zero, $p < .01$.

from the full regression models were tested against regression results from models restricted to biographical and job-related variables (B). As shown by the F test results in Table 2, the removal of the OAI item set from the regression equations was statistically significant for all criterion data sets--supporting the second hypothesis concerning OAI relationships with the criteria in the presence of the biographical and job-related control variables. These results were interpreted as providing substantive evidence for the linkage between the OAI and global job satisfaction, reenlistment intent, and reenlistment behavior. The last column of Table 2 presents the squared multiple correlations for the OAI equations (C). All coefficients were significantly different from zero. Prediction of reenlistment behavior appeared to be somewhat greater in magnitude for eligible airmen than for the eligible-ineligibile classification. As would be expected, the concurrent validations of the OAI against attitudes of global job satisfaction and reenlistment intent were somewhat higher than the predictive validations against reenlistment behavior, both for 1973 and for 1975 survey samples.

Specific Occupational Attitude Item Contributions

The third hypothesis proposed that the attitudinal saliency of specific OAI items would be similar across years for the separate criterion measures. The 189 OAI items were consecutively entered into multiple regression equations using a stepwise technique. Results from the final stepwise equations were examined to determine the relative predictive efficacy of individual items. The zero order correlations and the order of entry for the five most significantl predictive items associated with the criteria are presented in Table 3. Since prior findings indicate that the eligible only classifications produce higher correlations, the results are limited to the eligible only data sets.

Table 3
Criterion Correlations and Order of Entry in Stepwise Regressions for the Top Five OAI Items

Specific Job Satisfaction Item from OAI	Global Job Satisfaction 1973	Global Job Satisfaction 1975	Reenlistment Intent 1973	Reenlistment Intent 1975	Retention 1973	Retention 1975
Challenge provided by job	.73(1)	.65(1)				
Way job uses abilities	.73(2)	.63(2)				
Accomplishment feelings	.71(3)	.63(4)				
Amount of interesting work		.66(3)				
Supv brings out best	.41(4)					
Pace of your work	.49(5)	.45(5)				
Pay compared with outside			.35(5)	.35(1)	.20(4)	.22(1)
Consideration given by A.F.			.38(1)	.33(2)	.23(2)	.20(2)
Benefits compared to outside			.38(2)	.33(3)		.20(3)
Social position in A.F.					.24(1)	
A.F. removes irritants			.36(3)	.31(5)		
Contribution to nat'l defense			.32(4)	.26(4)		
Information on promotions					-.02(3)	
Recreation in community						.20(4)
Weighted Amn Promotion System						.18(5)
Educational opportunities					.14(5)	
Data sets (eligible only) N=	961	3,753	961	3,753	835	2,988

Three major inferences may be drawn from an inspection of the relationships displayed in Table 3. First, global job satisfaction appears to be associated with a different domain of specific occupational attitudes than reenlistment intent and behavior. Challenge, use of abilities, accomplishment feelings, and the pace of the work are common to both 1973 and 1975. Secondly, reenlistment intent and actual reenlistment behavior appear to be aligned on three items across all years, viz., pay and benefits, compared to civilian jobs, and the consideration given airmen by the Air Force. Two other items align with reenlistment intent across both years were the removal of irritants and contributions to the national defense. Airmen indicating low attitude scores on these items are more likely to separate than airmen indicating that they are satisfied with these issues. Finally, actual reenlistment behavior also appears to be influenced by social position and educational opportunities in 1973, shifting toward recreation and promotion concerns in 1975.

Cross-validation of the OAI Equations

The final phase of this research project examined hypothesis 4 concerning the consistency of the regression equations. Dual cross-validations were conducted upon all criterion data sets by the application of the 1973 least squares regression weights to the 1975 data sets, and vice versa. The resulting coefficients were then assessed against a correlation of zero to determine if the weights associated with the specific regression equations were stable enough to produce an acceptable level of prediction in another sample. Results from all F tests were significant ($p < .01$), and indicated that the specific occupational attitude effects were consistent across time, and that the multiple relationships observed in the development samples were not entirely attributable to the capitalization upon specific sample variance, but rather were indicative of stable patterns which could be replicated in other samples.

CONCLUSIONS

The accomplishment of the validation of the OAI with respect to global job satisfaction, reenlistment intent, and reenlistment behavior resulted in the following conclusions.

1. The OAI possessed concurrent validity against expressions of global job satisfaction and reenlistment intent and predictive validity with respect to actual reenlistment behavior. This validity was demonstrated for the OAI with and without the consideration of baseline effects due to number of biographical and job-related factors. Cross-validation results further indicated that these findings were generalizable across time.

2. The major areas of OAI attitudes identified as having important influences on global job satisfaction were: the challenge provided by one's job, the way the job uses one's abilities, the amount of interesting work one does, the feelings of accomplishment one gets from the work, and the pace of work.

3. The major areas of OAI attitudes identified as having important influences on reenlistment intentions were: pay compared with a civilian job, consideration given by the Air Force, fringe benefits compared with a civilian job, Air Force efforts to remove irritants, and contributions to the national defense.

4. The major areas of OAI attitudes identified as having important influences on actual reenlistment rates were: pay compared with a civilian job, consideration provided by the Air Force, fringe benefits compared with a civilian job, social position, promotion concerns, and recreational and educational opportunities.

5. While the OAI was successfully validated against global job satisfaction, reenlistment intent, and actual reenlistment behavior, it was found that the most highly related CAI items were different for global job satisfaction as compared with reenlistment intent and reenlistment behavior. Thus, is should be understood that enhancement of global job satisfaction may not necessarily bring about improvements in reenlistment intent and actual reenlistment. Likewise, successful efforts in relation to increasing reenlistment may not impact global job satisfaction concerns of first-term enlis..d airmen.

6. The findings of this study emphasize that post-enlistment occupational attitudes are important in addition to selection, classification, and assignment concerns related to global job satisfaction and reenlistment intent. More importantly, specific areas were shown to influence actual reenlistment rates. Efforts to enhance retention and limit separations in the first term could be realized by integrating these specific issues with curriculum materials for the career advisor, NCO Academy, and enlisted supervisor and management training programs.

References

Bottenberg, R.A., & Ward, J.H., Jr. Applied multiple linear regression. PRL-TDR-63-6, AD-413 128. Lackland AFB, TX: 6570th Personnel Research Laboratory, Aerospace Medical Division, March 1963.

Edwards, J.O., Jr. Comparative analyses of enlisted job satisfaction as measured by the Occupational Attitude Inventory. AFHRL-TR-78-61, AD-A063 642. Brooks AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory, October 1978.

Finstuen, K. An open role system perspective in analyzing self and social job attitudes. (Doctoral dissertation, The University of Texas at Austin, 1981). Dissertation Abstracts International, 1981, in press.

Gould, R.B. Review of an Air Force job satisfaction research project: Status report through September 1976. AFHRL-TR-76-75, AD-A035 684. Lackland AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory, December 1976.

Gould, R.B. Air Force Occupational Attitude Inventory development. AFHRL-TR-78-60, AD-A062 987. Brooks AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory, October 1978.

Pritchard, R.D., & Shaw, J.B. Comparison of published measures of job satisfaction on a taxonomy of job rewards. AFHRL-TR-78-21, AD-A058 138. Brooks AFB, TX: Occupation and Manpower Research Division, Air Force Human Resources Laboratory, July 1978.

Tuttle, T.C., Gould, R.B., & Hazel, J.T. Dimensions of job satisfaction: Initial development of the Air Force Occupational Attitude Inventory. AFHRL-TR-75-1, AD-A014 796. Lackland AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory, June 1975.

Tuttle, T.C., Hazel, J.T. Review and implications of job satisfaction and work motivation theories for Air Force research. AFHRL-TR-73-56, AD-782 443. Lackland AFB, TX: Occupational Research Division, Air Force Human Resources Laboratory, January 1974.

Ward, J.H., Jr., & Jennings, E. Introduction to linear models. Englewood Cliffs, N.J.: Prentice Hall, 1973.

Freda, Jon S., Hall, Eugene R. & Ford, Larry H., US Navy Training Analysis and Evaluation Group, Orlando, Florida.

Differential Effectiveness and Efficiency of Individualized Instruction: II. Major Findings

This paper presents the major findings of the TAEG differential effectiveness and efficiency study. Multiple regression analyses indicated significant differences in fleet supervisor ratings for graduates of individualized versus conventional instruction. These differences were related to different kinds of training tasks, but not to ability levels of graduates. Significant interactions between method of instruction and type of task, and method of instruction and ability level were found with respect to school achievement (for both course completion times and final course grades). The findings are discussed in terms of their utility for instructional design.

DIFFERENTIAL EFFECTIVENESS AND EFFICIENCY OF INDIVIDUALIZED INSTRUCTION: II. MAJOR FINDINGS

Jon S. Freda, Eugene R. Hall, and Larry H. Ford

Training Analysis and Evaluation Group
Orlando, Florida 32813

Individualized instruction has become a controversial issue in military training. Many individuals in both training and operational settings have come to believe that individualized instruction (II) is not a desirable or effective way to train students for operational job assignments. The widespread belief is that conventional classroom group-paced (GP) methods result in better trained personnel.

Currently, the U.S. Navy conducts technical training under both II and GP instructional methods. The principal II methods used are computer managed instruction (CMI) and self-paced (SP), or instructor managed instruction (IMI). Because of the potential for reduced student training time, the Navy plans to individualize still more of its courses. However, in view of concerns expressed by fleet units, the Chief of Naval Education and Training (CNET) tasked the Training Analysis and Evaluation Group (TAEG) to conduct a study to examine the effects of individualized instruction.

PURPOSE

The purpose of the study was to determine if individualized instruction is more or less effective and/or efficient than conventional instruction, and further, if these effects differentially relate to:

- training individuals of differing ability levels and/or
- training different types of tasks.

The present paper presents selected major findings of this study. The previous paper in this volume (Hall and Freda) presents details of the methodology employed to conduct this study.

ANALYTICAL STRATEGY

A partial hierarchial regression model was employed to examine the effects of each set of predictors on the criterion variables (Cohen & Cohen, 1975; Kim & Kohout, 1975). This model allowed a unique partitioning of the total variance of each criterion to be accounted for by each subset of predictors entered into the regression equation. The use of a multiple regression approach reflects current methodological approaches used in investigating aptitude-treatment interactions (Cronbach & Snow, 1977). Predictor variables were considered statistically significant and relevant if (1) they met the acceptable level of significance ($p < .05$) and (2) could explain a relevant amount of variance on the criterion variable (increment the multiple R^2 by at least 2 percent).

RESULTS

The results are presented first in order of the training efficiency and training effectiveness measures. Then, within each of these measures, significant predictors (both main effects and interactions) are delineated for each criterion variable. The results presented below are based on a data analysis across courses (see table 1). Only statistically significant results are presented.

Table 1.
Original Sample Size of Graduates by Method of Instruction

Method of Instruction	N
Self-Paced (SP)	1487
Computer Managed Instruction (CMI)	823
Individualized (SP + CMI)	2310
Conventional (Group-Paced (GP))	1696
Total	4006

Training Efficiency Measures

Criterion: Time to Complete the Course

Method of Instruction. II (SP + CMI) graduates completed their courses in a shorter period of time than CI (GP) graduates. CMI graduates completed their courses in a shorter period of time than SP graduates (see table 2).

Table 2.
Mean Time to Complete the Course (Contact Hours)

Method of Instruction	Mean	S.D.[1]	S.E.M.[2]	N
SP	166.35	69.60	1.94	1283
CMI	139.57	88.25	3.11	803
II (SP+CMI)	156.04	82.57	1.81	2086
CI (GP)	195.22	122.05	2.95	1707

[1]S.D. = Standard Deviation
[2]S.E.M. = Standard Error of the Mean

Ability Level. In general, graduates with higher AFQT percentiles finished their courses in a shorter period of time than those with lower AFQT percentiles (see table 3).

Table 3.
Mean Time to Complete the Course (Contact Hours)

AFQT Percentile Range	Mental Category	Mean	S.D.	S.E.M.	N
93-99	1	134.64	93.05	6.53	203
65-92	2	178.19	113.29	3.39	1114
49-64	3U	174.11	103.45	2.63	1551
31-48	3L	172.12	91.30	3.23	800
21-30	4U	191.73	83.40	14.30	34
10-20	4L	480.00	0.0	0.0	1

Method of Instruction by Ability Level. II graduates in the upper mental categories finished their courses in a shorter period of time than CI graduates in the upper mental categories. Both II (SP + CMI) and CI (GP) graduates in the mid and lower mental categories took about the same amount of time to complete their courses (see figure 1).

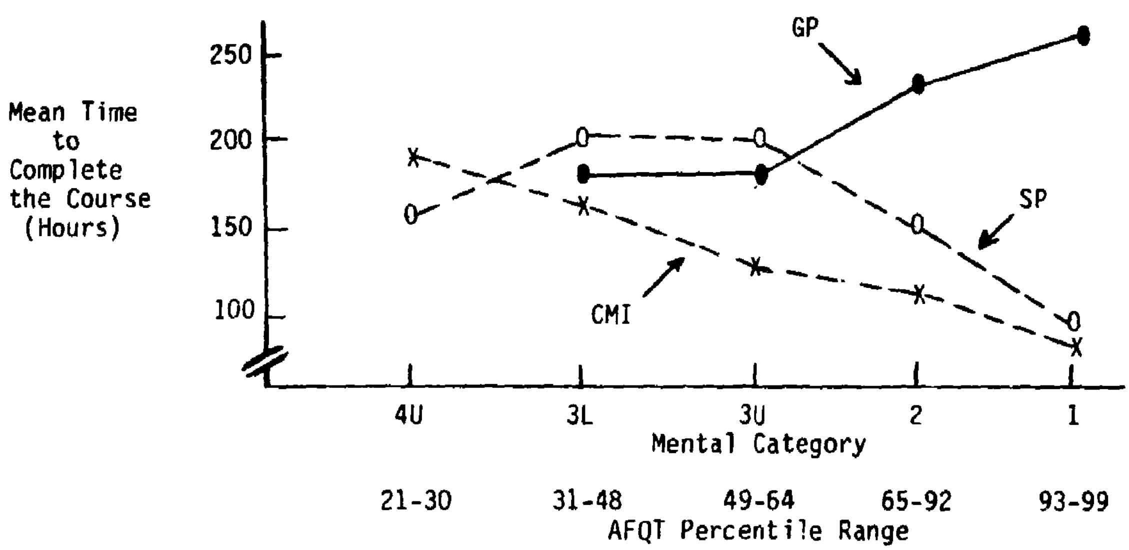


Figure 1. Mean Time to Complete the Course by Method of Instruction and Ability Level (i.e., Mental Category/AFQT Percentile Range). Mean Data Points Based on Less Than Five Graduates Are Not Plotted.

Training Task. In general, more fact tasks were taught in courses associated with longer completion times than with shorter completion times. More category, procedure, and rule tasks were taught in courses associated with shorter completion times than with longer completion times.

Method of Instruction by Training Task. II graduates had shorter completion times than CI graduates in courses that taught a smaller percentage of fact, category, procedure, and rule tasks. Both II and CI graduates took about the same amount of time to complete courses that taught a greater percentage of these tasks, excluding fact tasks (see figure 2).

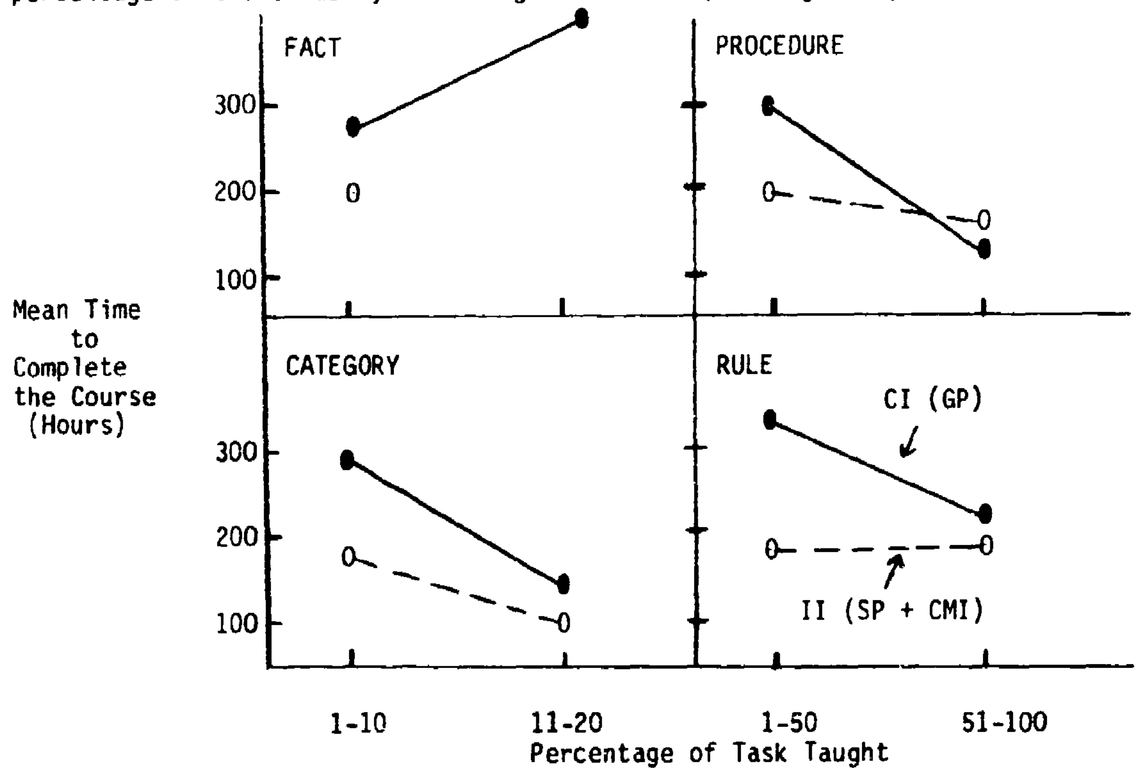


Figure 2. Mean Time to Complete the Course by Method of Instruction and Type of Task. There Is Only One Mean Data Point for II Fact Task

Criterion: Training Costs

Training costs refer to the total training costs to produce one graduate per course session. Training costs differ by each school, and all the graduates per school are assigned the same value of total training costs for that particular school. Thus, total training costs is a course level variable because each student does not have assigned to him/her a unique training cost . Rather, the training costs are assigned on a school (course) level. Therefore, training costs was not entered into the multiple regression model as a criterion variable due to the lack of variance in training costs within courses. Regression analyses of derived individual level training costs, as well as aggregated course level training costs, will be contained in reports currently being prepared by the TAEG.

Training Effectiveness Measures

Criterion: End of Course Grades

Method of Instruction. II (SP + CMI) graduates received higher end of course grades than CI (GP) graduates. SP graduates received higher end of course grades than CMI graduates (see table 4).

Table 4.
Mean End of Course Grade by Method of Instruction

Method of Instruction	Mean	S.D.	S.E.M.	N
SP	90.84	5.51	.23	584
CMI	85.74	8.24	.31	692
II (SP+CMI)	88.08	7.56	.21	1276
CI (GP)	82.10	6.72	.24	768

Ability Level. In general, graduates with higher AFQT percentiles received higher end of course grades than those with lower AFQT percentiles (see table 5).

Table 5.
Mean End of Course Grade by AFQT Percentile Range and Mental Category

AFQT Percentile Range	Mental Category	Mean	S.D.	S.E.M.	N
93-99	1	92.78	6.51	.56	135
65-92	2	87.45	7.63	.29	673
49-64	3U	84.55	7.32	.27	737
31-48	3L	83.58	7.63	.37	436
21-30	4U	81.87	7.83	1.71	21
10-20	4L	79.50	0.0	0.0	1

Method of Instruction by Ability Level. II (SP + CMI) graduates with higher AFQT percentiles received higher end of course grades than CI (GP) graduates with higher AFQT percentiles. Both II and CI graduates with lower AFQT percentiles received similar end of course grades. CI graduates with higher AFQT percentiles received similar end of course grades as CI graduates with lower AFQT percentiles. For both SP and CMI graduates, AFQT was positively related to end of course grade. There were no significant differences between SP and CMI graduates on end of course grades by ability level (see figure 3).

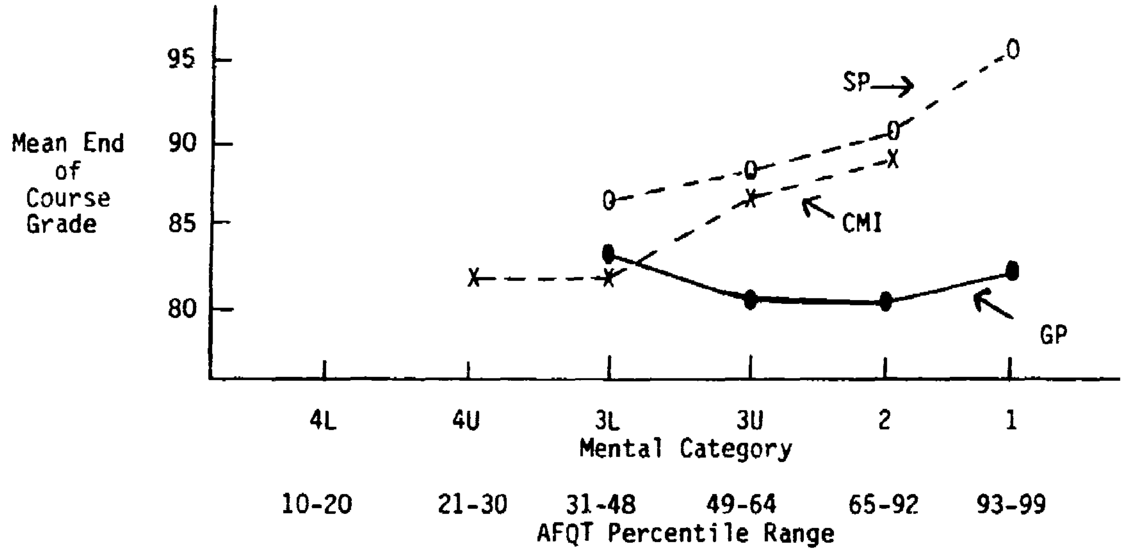


Figure 3. Mean End of Course Grade by Method of Instruction and Ability Level (Mental Category/AFQT Percentile Range). Mean Data Points With Less Than Five Graduates Are Not Plotted.

Training Task. In general, graduates received higher grades in courses that taught more category, procedure, and rule tasks, and received lower grades in courses that taught more fact tasks.

Method of Instruction by Training Task. II graduates received higher grades than CI graduates in courses that taught a smaller percentage of fact and category tasks. This difference between II and CI graduates is accentuated even more in courses that teach a larger percentage of category tasks (see figure 4).

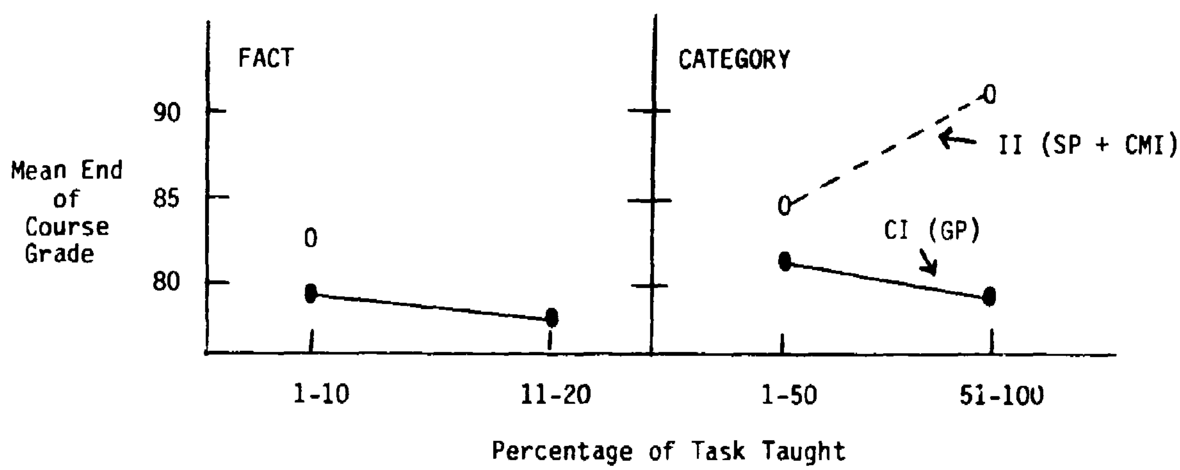


Figure 4. Mean End of Course Grade by Method of Instruction and Type of Training Task. There Is Only One Mean Data Point for II Fact Tasks.

Criterion: Training Appraisal System (TAS) Ratings

Method of Instruction. II (SP + CMI) graduates received higher TAS ratings than CI (GP) graduates. There were no significant differences between SP and CMI graduates on TAS ratings (see table 6).

Table 6.
Mean TAS Ratings by Method of Instruction

Method of Instruction	Mean	S.D.	S.E.M.	N
SP	2.90	.70	.02	899
CMI	2.96	.57	.03	287
II (SP+CMI)	2.91	.67	.02	1186
CI (GP)	2.69	.67	.02	1129

Training Task. In general, graduates who attended courses that taught a smaller percentage (less than 10 percent) of fact and category tasks received higher (mean = 2.90) TAS ratings than graduates who were taught a greater percentage (20-30 percent) of these tasks (mean = 2.70).

DISCUSSION

II graduates completed their courses in a shorter period of time and received higher end of course grades and training adequacy ratings than CI graduates in this study. These differences between II and CI graduates are also related to different kinds of training tasks and ability levels of the graduates. These results are discussed in terms of school and fleet performance measures. School performance criterion measures (end of course grades, time to complete), revealed significant two-way interactions between method of instruction, and ability level and training task. In terms of training efficiency, higher mental category II graduates completed their courses in less time than higher mental category CI graduates. There were no significant differences between lower mental category II and CI graduates in terms of time to complete. With respect to training effectiveness, II graduates with higher ability levels received higher end of course grades than CI graduates with higher ability levels. Both II and CI graduates with lower ability levels received similar end of course grades.

The relationship observed for the II graduates agrees with the generally reported finding of general ability measures predicting learning of new material (Cronbach and Snow, 1977). However, the relationship found for the CI graduates requires reflection as to its source. It is possible that CI courses with longer durations unintentionally selected higher mental category students than CI courses with shorter duration. This selection could have come about by the ASVAB composite entrance requirements for each CI course being related to AFQT percentiles.

An alternative explanation is that CI provides a less demanding and/or less controllable environment for the higher ability students. It has been suggestd that learning depends on general intellectual dvelopment to a greater degree when active intellectual work is required of the student. Methods of instruction that reduce the intellectual demand often reduce the differences between high and low ability students (Cronbach and Snow, 1977). If these methods are applied to instruction of low ability students over a long period, many low ability students may equal or excel high ability students in terms of their mastery of lesson content. This explanation is supported by the facts that (1) CI graduates cannot control the length of stay within the course, (2) CI graduates received similar end of course grades regardless of mental category, and (3) CI graduates obtained lower end of course grades when taught a greater percentage of complex training tasks than II graduates.

For fleet performance criterion measures (TAS ratings), ability level of the graduates was not related to the training adequacy rating. Overall, graduates who were taught a smaller percentage of fact and category tasks received higher TAS ratings than those taught a greater percentage of these tasks. It is noted that this latter finding is similar to the relationship of fact tasks with end of course grades, but opposite to the relationship of category tasks with end of course grades. Apparently, percentage of fact tasks taught in II and CI courses acts as a reliable indicator of end of course grades and TAS ratings; whereas, the influence of category tasks may be more susceptible to interfering factors during training and/or during the time interval between school graduation and fleet performance rating.

It is also noted that the sex of graduate and geographic location of the school were not significantly related to school or fleet performance measures.

Based on these results, it is suggested that higher ability students could be tracked into a more individualized instructional environment and lower ability students be tracked into a more conventional instructional environment. This suggestion assumes that an experimental setting could be designed in which a specific course would have conventional (instructor-group oriented) instruction aspects and individualized (self-pacing) instruction aspects.

It is also suggested that the "method of instruction" distinction could, then, be dropped in favor of a training task classification. This training task classification could be used, for example, to select incoming students for the individualized or conventional aspects of the course based on ability-task entrance scores.

CONCLUSIONS

Individualized instruction (II) compared favorably to conventional instruction (CI) with respect to the school and fleet performance measures used in this study. Ability level of graduates and type of training task significantly differentiated the school performance differences between II and CI, but did not interact to facilitate explanation of the differences between II and CI on fleet performance measures. There were no significant school or fleet performance differences due to different types of training tasks presented to different ability levels of graduates. The advantage of classifying the methods of instruction into different types of training tasks, then, appears to be one of clarifying the relative amounts of different kinds of information taught in the courses. Also, ability level of graduates was not related to fleet performance measures. Thus, the traditional factors described in the aptitude/treatment interaction literature (i.e., ability level, training task, and method of instruction) were shown to have more predictive power (i.e., larger number of significant main effects and interactions) for school performance measures, but to have less predictive power for fleet performance measures in this study.

REFERENCES

Cohen, J. and Cohen, P. Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers, 1975.

Cronbach, L. J. and Snow, R. E. Aptitudes and instructional methods. New York: Irvington Publishers, Inc., 1977.

Kim, J. and Kohout, F. J. Multiple regression analysis: subprogram regression. In N. Hie, C. Hull, J. Jenkins, K. Steinbrenner, and D. Bent (Eds.), Statistical package for the social sciences (2nd ed.). New York: McGraw-Hill Book Company, 1975.

AD P001317

An Enriched User Oriented Cost and Training Effectiveness Analysis Methodology

Edward W. Frederickson, Ph.D.
Principal Scientist
Applied Science Associates, Inc.
El Paso, Texas

Applied Science Associates, Inc. has developed and applied a CTEA methodology that can be used with developing, as well as already fielded, equipment systems. The methodology is grounded in a thorough front-end-analysis of the job and training requirements. The products of these analyses feed not only the cost and effectiveness trade-off analysis but provide a significant portion of the structure and content for actual course design. Trainability assessment and training alternative selection become less significant issues when compared to the usefulness of the products that are relevant to course development.

Introduction

The El Paso Office of Applied Science Associates, Inc., has developed an approach to the conduct of a Cost and Training Effectiveness Analysis that emphasizes a thorough analysis of a job and its tasks to specify the functional learning requirements of the training system. The procedures were developed for formalizing the collection, analysis, and integration of training systems cost, impact and effectiveness data. The primary purpose of the analysis and integration of training-related data concerns the comparison of the cost and effectiveness of alternative training programs for meeting pre-defined performance objectives. The secondary purpose of the analysis is the determination of the trainability of the job performance requirements under the specific training constraints and conditions.

These two purposes, cost effectiveness and trainability, are common to most CTEA methodologies. ASA's approach to fulfilling these purposes differs somewhat in the assessment of training effectiveness, and the determination of trainability. The major procedural differences lie in the use of a technique for quantifying training effectiveness, and in the use of an extended program of instruction (POI) for determining trainability. These techniques will be briefly covered later, but first the primary purpost of this paper will be addressed--the development of products for users other than the traditional user of CTEA results.

The basis of the ASA approach is a human factors oriented front end analysis (FEA) that is conducted with the following additional purposes in mind:

1. Training requirements
2. Training system design
3. Training device specification
4. Training management
5. Training course design
6. Training system design (and career development)
7. Training performance assessment
8. Training support plan to include support products

<u>Background</u>

These issues are addressed as a necessary part of the complete CTEA as carried out following ASA procedures. ASA's CTEA philosophy is that the validity and completeness of the analytical process is dependent upon the FEA. If the FEA is thorough, the resulting CTEA decisions objectively fall out of the analysis and relatively few instructional delivery system decisions have to be made. Using the augmented CTEA procedures for two Army systems we found that our analytical process yielded outputs beyond those usually expected from a CTEA. The procedures emphasize a thorough, detailed FEA that provides data and information of use by other than the usual CTEA users, such as course designers, course developers, the training device system, career development planners, performance evaluation developers, training managers, course instructors and training support planners.

CTEA requirements were instituted with the introduction of the Life Cycle System Management Model which guides and controls the conceptualization, development and deployment of major materiel systems. As initially conceptualized, the CTEA procedures addressed the two major purposes. First, the analysis concerned the comparison of the cost and effectiveness of alternative training programs so that a program could be selected for meeting defined performance objectives. ASA's CTEA, as do most other approaches, assumed that the performance objectives, or at least task statements, would have been previously prepared and available for analysis. This assumption proved to be not entirely correct and job analysis procedures had to be added to the CTEA process. The second concern was with the determination of the trainability of the job performance requirements under the specific training conditions. It became obvious as we got into the two CTEA projects that issues in addition to the training conditions impacted trainability in a way so as to constrain the degrees of freedom in program design and delivery options. As a result, a procedure was adopted to identify the constraining factors as early as possible in a CTEA project. Doing this keeps the analyst from wasting time and effort in the pursuit of impossible training alternatives.

As we reviewed the procedures for conducting the job analysis and trainability determination, it was decided that the best way to carry out the several analyses leading to the evaluation of training alternatives was to use as complete a description of the training requirements as possible. To present these requirements we adopted a program of instruction (POI) format. Actually, more information was to be presented than is usual so the term "extended POI" was adopted. The information presented in the extended POI was generated through several sets of analytical procedures beginning with the specification of tasks. The sets of procedures were designed to meet several process objectives, such as standardization of procedures, reduction of requirements for personnel time and resources, provision of an objective audit trail, quantification of the training alternatives comparisons and establishment and management of a reusable data base.

ASA's CTEA Process

I want to stress that ASA's CTEA process accomplishes the purposes required in the LCSMM, as does most other approaches. However, ASA's approach provides additional products from the analytical procedures that result in a more integrated and resource efficient training development system.

The various blocks of procedures used in our approach will now be briefly presented. There are ten major blocks of activities as follows:

1. Preparation of work plan
2. Analysis of missions and functions
3. Selection of tasks for training
4. Analysis of tasks
5. Generation of general course structure
6. Generation of training program alternatives
7. Development of extended POI
8. Analysis of training effectiveness and trainability
9. Analysis of training costs
10. Final trade-off analysis

The first seven sets of activities are preliminary, but absolutely essential, to the actual conduct of the cost and effectiveness analyses.

The methodology represents an integration of various elements of the state-of-the-art in instructional technology, course design, and collective and individual front-end analyses. The first seven sets of activities are described as preliminary in that they produced the various assumptions, constraints, training development materials, decision data and information required for the generation of training effectiveness estimates and the determination of cost differences for the specified training program options. These seven sets of activities are the key to the ASA CTEA methodology. The rationale for our emphasis on the FEA approach draws upon the purpose of CTEA and its relationship to the materiel system development process.

The specific nature of a CTEA often is dependent upon the state of development of the materiel system under study. For conceptual materiel systems, the lack of performance data will mean that CTEA primarily will be used to forecast training resource requirements and to indicate potential problem areas in the training program. In other words, a CTEA begun early in the materiel development process will identify training issues that may require special examination during later developmental and operational testing (DT/OT). As prototypes of the materiel system become available, CTEA would involve updating and validating cost and resource impact projections, and analytical investigations of training effectiveness. Following the field deployment of the materiel system, the emphasis of CTEA would be on the cost effectiveness of (1) training "fixes" designed to address training deficiencies, or (2) training modifications designed to meet an altered threat scenario or to accommodate evolutionary hardware modifications.

Preparation of Work Plan

Once the status of the LCSMM for the materiel system has been determined, all relevant documentation for the system, antecedent and similar systems are identified and located. This is a very important step for the CTEA in that the documentation provides a definitive basis from which to launch the study. As the LCSMM process proceeds, additional documentation and changes to the initial set of materials must be obtained. This material is assembled, cataloged for future CTEA activities, and reviewed. The initial and subsequent sets of documents becomes the first element of the audit trail. They also provide the beginning list of constraints that must be dealt with.

The location of equipment and personnel familiar with the equipment are next determined. This is necessary in order to prepare the detailed work plan. Interviews with these subject matter experts (SMEs) are scheduled as required in the CTEA process. The detailed plan generally follows the CTEA block diagram of major activities. Forty nine specific steps are described in the detailed plan to reflect the support requirements, time schedules, data analyses, and output products.

Mission and Function Analysis

Concurrently, work can begin on the specification of the initial job task list. This begins with a review of the documentation to identify the materiel system missions. Missions are analyzed into functions, which are major chunks of mission activities usually assigned to specific job positions. Where job descriptions already exist and MOSs have been designated for the positions, the mission analysis is often not necessary.

ASA uses systematic sets of procedures for analyzing categories of functions. The functions are analyzed with the assistance of SMEs who provide specific responses to questions which pertain to the function being analyzed. SMEs identify specific actions performed on specific objects. These responses are subsequently written as task statements. This analysis process produces the initial set of tasks, which is the second audit trail item.

Selection of Tasks for Training

Since training time is limited, all job tasks usually cannot be included in training. Using guidelines from the ISD process, ASA developed a set of procedures that first selects tasks that should be trained. Then it specifies the level of training that is required. This information has two CTEA benefits: (1) A rational, objective basis is provided for selecting tasks for training; and (2) Information is generated for designing the entire training system for the MOS, to include resident, unit, refresher, maintenance, and wartime preparatory training requirements. The information for making these decisions comes from the analysis of the criticality of each task. Less critical tasks require only the use of simple procedures or general skills and thus may not need to be trained. Remaining tasks can be ordered in terms of the ten criticality dimensions that are used:

1. Learning difficulty
2. Performance difficulty
3. Time delay tolerance
4. Consequences of inadequate performance
5. Immediacy of performance
6. Civilian acquired skills
7. Task importance
8. Frequency of performance
9. Wartime task
10. Task decay rate

Each item is assessed on a three level scale by SMEs. The responses are then subjected to logical analysis using a computer routine. Each task is classified into one of the following categories:

1. Requires certification training
2. Requires qualification training
3. Requires wartime refresher training
4. Can be considered for on the job training
5. Requires maintenance of proficiency training
6. Can be considered for reduction of training time
7. Can be considered for elimination from training

Analysis of Tasks

The tasks that are selected for training are next reviewed by SMEs who provide task content and context information. Four sets of information are obtained--performance standards; unusual job conditions; stimuli that must be attended to during task performance; and the skills required for performing the task. The products of this analysis are the performance objectives which become the terminal learning objectives, the functional learning requirements and the instructional delivery systems that are appropriate for each task.

Generation of General Course Structure

This analysis requires SMEs to provide inter-task dependency information for each task. SMEs identify several tasks that they must know how to perform before they can carry out a specified task. The entire list of tasks selected for training is reviewed and 3-5 subordinate tasks are identified for most of the tasks. There are always several tasks at the bottom of the hierarchy where only one or two subordinate tasks are designated, and one or two that fall at the very bottom so have no subordinate tasks. The hierarchy information is subjected to analysis by a computer routine which provides a printout presenting the relationships between all tasks and skills.

ASA's approach to the structuring of training programs stems from our job structure philosophy--that job tasks are accomplished by the application of skills and the carrying out of simple procedures. Simple procedures can be carried out by merely following written or verbal instructions. Special and general skills are acquired through practice and thus must be trained. General skills are usually already in the behavior repertoire of the general student population and usually do not need to be trained. Special skills thus are the primary substance of training programs. Skills are applied across work tasks so should be taught in the same manner. In ASA's CTEA, the general course structure is determined through an iterative process beginning with the integration of the task and skill hierarchies. Tasks that require skills, which fall at the bottom of the skill hierarchy, are placed early in the course structure. As you move through the course structure, more new skills are introduced. Tasks placed later may include both skills that have already been taught or have not yet been introduced. The "new" skills are then trained, and the "old" skills already taught, are practiced.

The course structure is modified from the ideal according to the specific constraints that apply at the proponent school. Constraints are issues that impact the design, delivery and management of a course. These issues are identified throughout the CTEA study in interviews with school personnel, TRADOC Systems Management (TSM) personnel, and in the reviews of antecedent or similar systems training materials. By the time that the general course structure is generated almost all of the constraints have been identified. These include the following:

1. Available time and resources
2. Instructors--number and qualifications
3. Facilities
4. Management structure
5. Student population
6. Student load
7. Course content
8. Instructional method philosophy
9. Media and training devices available

Generation of Training Program Alternatives

Once a general course structure has been developed, the next step is to specify training program alternatives which could be used to accomplish the training objectives. The variables that can be manipulated include the instructional delivery system, the training time (peacetime, shortened peacetime, mobilization time), logistical support, types of simulators, student characteristics and numbers, facilities, and numbers of instructors. The actual selection of training alternatives does not involve a free choice of alternative variables or levels of variables. School philosophies, doctrine, and local SOPs, in addition to any training structure determined during the materiel system conceptualization, narrow the list of possible alternatives to a relatively small list.

Two approaches can be used to select alternatives. This first is to identify two extreme levels of a few variables that would have the greatest impact on cost and effectiveness estimates and then propose the two extreme combinations as the alternatives. The second approach would be to select alternatives that would be most acceptable to the school personnel, under their constrained conditions, and one or more ideal alternatives that essentially ignore the constraints. The ideal, if it proved to be the most cost/effective, would be an alternative that the school can plan subsequently to work towards.

Development of Extended POI

The last preliminary analytical block of activities is the development of an extended POI. The POI is the primary product of the first seven sets of CTEA activities. It is one document that provides almost all of the information needed for course development. It is the course design. It is designed so it can be used as it is, intact, or taken apart and recombined to meet individual course developer needs, bias or whatever.

Each page of the POI covers one task. The performance objective is written out. New and old skills are specified, which reflects the ordered structure of the entire POI. The instruction method and delivery medium are listed. And the instruction-to-practice time relationship is presented. An extended POI is prepared for each training program alternative, wherever necessary. Quite often different training program alternatives can use the same POI, since the dimensions determining differences are not all related to the instructional process. For example, where operational equipment is to be used that presents a display to the operator, various means can be used to generate the input signal. The actual stimulus source could be used, or any of a number of simulators could be used. As far as training the operator, however, it would make absolutely no difference how the display signal is generated, since the student would not see the stimulus source. But tremendous differences in costs could exist between the possible signal generators, each of which would be considered a different training program alternative.

Analysis of Training Effectiveness and Trainability

The POI is used as the primary stimulus for eliciting estimates of training effectiveness. ASA uses a forecasting approach for obtaining these estimates. Instructional experts are used as SMEs for estimating the effectiveness of training each task. The SMEs use the POI to provide the set for their estimates. The general instructions to the SMEs stress that each element of the POI must be reviewed and then kept in mind when estimating effectiveness. Two dependent variables can be used to estimate the utility of a training program--training time or training effectiveness. One of these is usually held constant, while the other is allowed to vary. Usually time is held constant and the effectiveness variable is manipulated by alternating training program variables. In the two Army CTEA projects, we did just that, held time constant and estimated effectiveness.

Actually the process is iterative in nature. Several questions are asked of SMEs. The POI is first presented without training time. Instructional design experts are asked to indicate how long it would take to train the task under normal conditions. They then indicate the minimum time for training under wartime conditions, and also provide a maximum time they would use if no time constraints are imposed.

The next step is to have SMEs estimate the effectiveness of training each task using all the information on the POI. In our Army projects, effectiveness was defined as the percentage of students that would reach the performance criterion under each time condition for each training alternative.

These effectiveness estimates are used to assess the trainability of each task. A task is trainable if the required percentage of students can reach the performance criterion level within the normal time frames to be established in the training program. By obtaining effectiveness estimates for each task, decisions can be made subsequently about program modifications for tasks that do not meet the trainability criterion.

Task criticality dimensions are used to provide a more quantitative basis for the comparison of total training program effectiveness. Tasks differ in terms of their training worth. ASA defines training worth in terms of five of the ten criticality dimensions used to select tasks for training. These five dimensions were scaled on the basis of the information utility they provide for determining the worth of including a task in a training program. These dimensions and their scale values are:

		Worth Value
1.	Consequences of inadequate performance	.45
2.	Task importance	.26
3.	Time delay tolerance	.16
4.	Frequency of performance	.08
5.	Immediacy of performance	.05

Each of the three levels of these dimensions are weighted. By multiplying the worth value by the response weight and by the effectiveness estimate, a figure of merit value is determined for each training program alternative. The results of this process are training effectiveness ratings for the programs, which are then compared by determining the relative effectiveness.

<u>Analysis of Training Costs</u>

The training cost analysis requires a rather detailed analysis of the training support requirements for each program. Specific requirements that are identified are numbers and types of support personnel, the hours they will be required during the entire course, equipment that must be dedicated or borrowed and hours it will be used, materials to directly or indirectly support the training, facilities to include classrooms, offices, maintenance shops, quarters, etc. The cost analysis approach that is used is often termed an incremental cost analysis. Since only relative cost comparisons are necessary to meet the CTEA purposes, all common costing elements are identified and eliminated from further consideration. Only those elements or levels of elements that are unique to a program are subsequently costed. Two kinds of costs are obtained; exact costs for unique equipment and materials; and average or scaled costs for personnel and common equipment items which are taken from cost data bases which are kept current by several agencies.

The total unique costs are determined for training program development, training program implementation and training program maintenance. These costs are then combined to provide a figure for a one year period. Next, life cycle costs for the training program are figured. These figures reflect only the cost of the unique training support requirements for each program. There is no intent here to provide an absolute cost estimate, since the purpose of a CTEA is to provide input to the decision of which training approach should be used. The last step is to determine relative program costs by forming ratios of cost figures, using one program as the base costing figure.

Final Trade-Off Analysis

The final CTEA step is to compare all program alternatives on both the relative effectiveness and relative cost figures. Since both effectiveness and cost data are relative figures and the cost ratios are based upon incremental cost figures, it can be somewhat misleading to combine the two sets of data into one figure of merit. A total figure would hide information that would be significant in selecting a program. Since both effectiveness and cost figures are relative, a tabular display is more appropriate as seen here.

Because of the structure of the POI and the detailed descriptions of the training support requirements for the different programs, modifications can readily be made to any program to alter the effectiveness and/or cost ratings. Any modification to the program would require subsequent changes to the POI and course structure, but this requires relatively little in the use of further resource expenditure as compared to the initial program design. In other words, fine tuning of a program to reduce costs or increase effectiveness is quite possible, bringing the realistic program towards an ideal cost/effectiveness level.

In summary, ASA's CTEA heavily emphasizes a detailed analysis of the job and the requirements for training that job prior to the actual analysis of the effectiveness and costs of alternative training programs. Information and data are generated through seven sets of preliminary activities that produce information that is of use to other users in the training development system.

Frey, Robert L., Jr., US Coast Guard Headquarters, Washington, DC.
(Wed. A.M.)

Prediction of Boot Camp Attrition: IRT vs. Number-Right Scoring

Item Response Theory (IRT) appears to be the wave of the near future in testing. Although the IRT literature is voluminous, it seems that there are, to date, virtually no data on the improvement in predictive validity from IRT scoring as compared to number-right scoring. The US Coast Guard use the Coast Guard Selection Test (CGST) for enlistment screening. The CGST is a battery of three tests: verbal ability, arithmetic ability, and mechanical comprehension. A large sample (ca. 2200) of Coast Guard enlistees was tracked through boot camp to obtain the criterion data on completion/attrition. For this sample, number-right scoring and IRT scoring were applied to each of the three subtests of the CGST. Comparisons were made on various indices of predictive validity between the two types of scoring. Implications of the findings for improving military selection procedures were discussed.

PREDICTION OF BOOT CAMP ATTRITION: IRT VS. NUMBER-RIGHT SCORING

Robert L. Frey, Jr.
USCG Headquarters

Karen Jones
USCG Institute

Item Response Theory (IRT) appears to be the wave of the near future in testing. Although the IRT literature is voluminous, the reported research comparing real-world predictive validity of IRT scoring vs. number-right scoring seems to be sparse. In the opinion of the authors, there is now a need for a number of such comparative studies in different organizations and settings. Only in this way will we obtain the necessary overview to detemine the practical utility of IRT scoring.

A detailed discussion of item response theory is not germane to the purpose of this study. However, Weiss and Davison (1981) present an overview of testing theories which includes a discussion of IRT research dealing with problems not resolvable witn classical testing theory. Suffice it to say that the major reason for our keen interest in the IRT scoring of conventional tests is the expectation that this procedure will result in improved predictive validity. In brief, we have this expectation because of the superiority of IRT scoring. IRT scoring provides estimates which are sample invariant and take into account each test item's characteristic curve. Consequently, IRT scoring results in a much more accurate estimate of one's ability level on the dimension in question (e.g., verbal, arithmetic) than number-right scoring. A logical inference, then, is that IRT scores should demonstrate better predictive validity than number-right scoring of the same test.

Specifically, this study compared number-right scoring vs. IRT scoring of the Coast Guard enlistment screening test in the prediction of boot camp attrition. The test battery used by the Coast Guard for enlistment screening is called the Coast Guard Selection Test (CGST). The CGST has three subtests: 1) General Classification Test (GCT), a verbal performance test; 2) Arithmetic (ARI); and 3) Mechanical Comprehension (MECH).

Method and Results

CGST subtest scores were obtained from a sample of 2138 Coast Guard recruits. Number-right scoring is the only method used for operational purposes. There were three number-right scores for each person: 1) GCT, 2) ARI, and 3) MECH. The raw scores were converted to Navy Standard Scores using equicentile calibration with the Navy BTB, Form 6. Navy Standard Scores have a theoretical mean of 50 and a standard deviation of 10 on an unrestricted population of service-age youth.

IRT Estimates of Examinee Ability

IRT estimates of examinees' ability on each of the three tests (GCT, ARI, and MECH) were obtained using estimates of item parameters and a Bayesian modal estimate scoring program. The latent trait or item response theory model selected to define the item characteristic curve or the regression of each binary-scored multiple choice item on latent ability was Birnbaum's (1968) logistic three parameter model. With this model, the item parameters are: item discrimination, a_i; item difficulty, b_i; and item coefficient of guessing or the lower asymptote, c_i (Hambleton and Cook, 1977; Urry, 1977).

The item parameters were estimated using the computer program ANCILLES. Since an earlier version of this program is discussed by Urry (1975), the discussion of the program in this paper will be limited to what is required to enable the reader to understand how the program was used. The item parameters were estimated for each of the tests separately. That is, the GCT, ARI, and MECH were analyzed separately as tests of 50, 35, and 25 items, respectively. ANCILLES provides the user with the option of the normal ogive or the logistic model. The logistic model was selected. The item parameters are estimated by an iterative minimum chi-square procedure in two stages and the user can obtain item parameters from either or both the stages. In the first stage, the distribution of ability is represented by corrected raw scores and in the second stage the distribution of ability is represented by Bayesian modal estimates of ability. The item parameters used in this investigation were from the second stage. When the item parameters were estimated, two items (one on GCT and one on ARI) did not fit the model and item parameters were not estimated by ANCILLES. Therefore, these items were omitted from computations in the next phase.

In this phase of the investigation, estimates of examinee ability were obtained using the item parameters provided by ANCILLES and a Bayesian modal estimate scoring program. Apparently, ANCILLES can be modified to provide estimates of examinee ability in addition to the estimates of item parameters. However, rather than modifying the program, a separate scoring program was used. Each test was again analyzed separately and, using the appropriate set of item parameters estimates, estimates of ability on the GCT, ARI, and MECH for each examinee were obtained.[1]

After the completion of recruit training, criterion data were obtained for each person - - that is whether or not the individual successfully completed recruit training. Before the predictive validity results are presented, a few descriptive statistics on the NSS scores and the IRT scores will be noted.

The NSS statistics were:

GCT - $\bar{X}$ = 55.08	S.D. = 8.49
ARI - $\bar{X}$ = 49.60	S.D. = 7.44
MECH- $\bar{X}$ = 50.37	S.D. = 7.82

The IRT statistics (theta estimates) were:

GCT - $\bar{X}$ = -0.02	S.D. = 0.83
ARI - $\bar{X}$ = 0.04	S.D. = 0.86
MECH- $\bar{X}$ = -0.06	S.D. = 0.72

The NSS GCT score correlated .9897 with the IRT GCT score. The NSS ARI score correlated .9786 with the IRT ARI score. The NSS MECH score correlated .9735 with the IRT MECH score.

In order to make interpretation easier, the IRT scores were then transformed to have the same means and standard deviations as the NSS scores. All subsequent findings for the IRT scores are based on this converted scale. In spite of the extremely high correlations between the IRT and NSS scores, however, the variance of the IRT scores at each NSS score level was not insignificant. The average standard deviation of the IRT GCT at each NSS level was 1.16. The average standard deviation of the IRT ARI scores at each NSS level was 1.40. The average standard deviation of the IRT MECH scores at each NSS level was 1.72.

[1]The authors wish to acknowledge Thomas A. Warm for his programming assistance which enabled them to use ANCILLES and the Bayesian modal estimate scoring program.

Even more noteworthy are the ranges of the IRT scores at each NSS level. The range of the IRT GCT scores is as large as 9 points at some of the NSS levels. The range of the IRT ARI scores is as large as 12 points at some of the NSS levels. The range of the IRT MECH scores is as large as 13 points at some of the NSS levels. In other words, at some NSS levels, people with the same number-right scores have IRT scores which differ by as much as 1.66 standard deviations of the entire sample. This result potentially allows for meaningful differences when comparing predictive validities.

A number of different analyses were done to compare the predictive validity of IRT scores vs. NSS scores. In the first analysis, the sample was stratified into groups based on their mental categories. The groups were: 1) CAT IV (AFQT score of 10-30), 2) CAT III B (AFQT score of 31-48), 3) CAT III A (AFQT score of 49-64), 4) CAT II and CAT I (AFQT score of 65-99). This was done on the basis of scores on the GCT+ARI composite. (The GCT+ARI composite is essentially a basic-skills achievement test.) Using the IRT scores, the results were as follows for the total sample:

	N	Attrition Rate
CAT IV	342	19.0%
CAT III B	457	10.1%
CAT III A	522	7.3%
CAT II + I	817	4.4%
Total	2138	8.7%

Using the NSS scores, the results were as follows for the total sample:

	N	Attrition Rate
CAT IV	335	19.4%
CAT III B	425	9.1%
CAT III A	509	8.5%
CAT II + I	842	4.3%
Total	2138	8.7%

As can be seen from the above tables, there was practically no difference in attrition rates using IRT scores vs. NSS scores. The IRT CAT III B and CAT III A groups do seem to show better differentiation in attrition rates (10.1% vs., 7.3%) than do the NSS groups (9.1% vs. 8.5%). Even if this result were to hold up on a new sample, there would be minimal practical significance, however.

In the next analysis, the group was again divided by mental categories. The GCT + ARI + MECH composite was used instead of the GCT + ARI composite. The results for the IRT scores were as follows:

	N	Attrition Rate
CAT IV	318	18.9%
CAT III B	536	11.6%
CAT III A	571	6.8%
CAT II + I	713	3.4%
Total	2138	8.7%

The results for the NSS scores were as follows:

	N	Attrition Rate
CAT IV	303	18.8%
CAT III B	543	12.0%
CAT III A	563	6.2%
CAT II + I	729	3.8%
Total	2138	8.7%

This time, the results are virtually the same for the IRT scores and the NSS scores. In this set of data, it seems that IRT scoring does not provide better predictive validity than number-right scoring when predicting boot camp attrition as a function of mental categories.

Next, discriminant function analyses were done. In brief, for the two group case, discriminant function analysis is a multivariate procedure which finds the one best linear combination of variables to enable the largest possible statistical separation between them. The procedure takes into account the vector of group mean differences and the correlations amongst the variables in determining the optimal weights for the linear combination. The hope is that a prediction equation will be found to make possible significantly more accuracy than is obtainable from knowledge of the base rates alone. In this case, the two groups are graduates (91.3%) and attritees (8.7%) and the variables are GCT, ARI, and MECH. For the IRT scores, the mean scores were as follows:

MEAN SCORES FOR EACH GROUP

GROUP	GCT	ARI	MECH
GRADUATES	55.5	50.0	50.7
ATTRITEES	50.6	45.8	46.4

As just mentioned, the major interest is in the accuracy of prediction possible based on the linear discriminant function derived from the analysis. The first classification table assumes equal prior probabilities for each group:

TRUE GROUP	PREDICTED GROUP (IRT SCORES): GRADUATE	ATTRITEE	TOTAL
GRADUATE	1240	713	1953
ATTRITEE	61	124	185
TOTAL	1301	837	2138

In this case, 1364 observations (63.8%) were classified correctly. Obviously, this is quite undesirable since a blind "hit" rate of 91.3% is possible simply by classifying everyone as a graduate.

The next table assumes prior probabilities proportional to sample size (91.3% and 8.7%)

TRUE GROUP	PREDICTED GROUP (IRT SCORES): GRADUATE	ATTRITEE	TOTAL
GRADUATE	1953	0	1953
ATTRITEE	184	1	185
TOTAL	2137	1	2138

In this case, 1954 cases were correctly classified. Unfortunately, the improvement from the blind "hit" rate is miniscule - one additional observation correctly classified.

Using the NSS scores, similar classification tables were generated. Assuming equal prior probabilities generated the following table:

TRUE GROUP	PREDICTED GROUP (NSS SCORES): GRADUATE	ATTRITEE	TOTAL
GRADUATE	1217	736	1953
ATTRITEE	62	123	185
TOTAL	1279	859	2138

In this case, 1340 observations (62.7%) were classified correctly. Once again, this is much worse than the blind "hit" rate.

The next table assumes prior probabilities proportional to sample size (91.3% and 8.7%).

TRUE GROUP	PREDICTED GROUP (NSS SCORES) GRADUATE	ATTRITEE	TOTAL
GRADUATE	1953	0	1953
ATTRITEE	185	0	185
TOTAL	2138	0	2138

Unfortunately, the results are exactly the same as the blind "hit" rate.

In summary, the discriminant function analyses showed little value for improving classification accuracy for both the IRT scores and the NSS scores. A major factor is that the proportion for the graduation group is so high (91.3%) that it is virtually impossible for any statistical technique to improve upon base rate prediction.

In the final statistical procedure, both graduation/attrition and ethnic group (minority/majority) were taken into account. In other words, a crossed factorial design was used. The two factors were Ethnic Group (G) and Attrition (A) -- a 2 X 2 factorial. Schematically, the four groups with the cell Ns were:

	Graduates	Attritees
Minority	269	42
Majority	1678	142

The attrition rate for minorities was 13.5% and the attrition rate for the majority was 7.8%. The three dependent variables once again were GCT, ARI, and MECH. Specifically, this procedure was a non-orthogonal multivariate analysis of variance. It can also be thought of as a combination of factorial design and discriminant function analysis. Using the IRT scores, the means for the design were:

	GCT	ARI	MECH
MINORITY GRADUATES	50.4	46.3	45.8
MINORITY ATTRITEES	46.2	42.4	41.9
MAJORITY GRADUATES	56.3	50.5	51.5
MAJORITY ATTRITEES	52.0	46.8	47.8

In this design, the test of the G X A interaction is a test for differential validity. That is, within each ethnic group, are the mean differences between graduates and attritees the same.

(Since the factorial design was non-orthogonal, the confounding effects of G and A were partialled out of the test for G X A.)

The multivariate F for the G X A interaction was 0.018, $p < .997$. This strongly indicates that there is no evidence whatsoever of differential validity. Of course this is virtually self-evident from the pattern of cell means.

Next, the test of the A factor (attrition) was done, partialling out the confounding effect of the G factor (ethnic group). The multivariate F for the A factor was 24.02, $p < .001$, $R = .181$. The standardized discriminant function coefficients and the discriminant loadings (or structure coefficients) for the A factor were:

	COEFFFICIENTS	LOADINGS
GCT	.385	.800
ARI	.406	.792
MECH	.485	.764

The coefficients are the weights applied to the standardized dependent variables to compute the discriminant scores. The loadings are the correlations of the dependent variables with the discriminant scores. The coefficients indicate the relative contribution of each variable to the statistical discrimination between the graduation and attrition groups. The loadings are used (as in factor analysis) to determine the underlying dimensionality of the discriminant function.

Since coefficients and loadings are subject to large sampling variability, any conclusions should be tentative, however. The coefficients indicate that the MECH test probably has a slightly larger relative contribution to distinguishing the graduates and attritees than do the GCT and ARI. The loadings seem to indicate that all three dimensions (verbal, arithmetic, mechanical comprehension) are required to define the discriminant function.

The test of the G factor (ethnic group) was done, partialling out the confounding effect of the A factor (attrition). The multivariate F for the G factor was 72.77, $p < .001$, $R = 305$. The coefficients and loadings for the G factor were:

	COEFFICIENTS	LOADINGS
GCT	.459	.791
ARI	.177	.650
MECH	.616	.847

The coefficients seem to indicate that the MECH test has the largest relative contribution to distinguishing between the minority group and the majority group, and further, that the ARI has almost no relative contribution.

Turning to the NSS scores, the same set of analyses were run - - G X A, G, A. The results were virtually the same. The multivariate F for the A factor (attrition) was 20.76, $p < .001$, $R = .169$. In other words, the difference in R values between the IRT scores and the NSS scores for attrition was .181 - .169 = .012. Needless to say, the diffference is miniscule.

Discussion

To review briefly, a number of comparisons were made between IRT scores and number-right scores as predictors of boot camp attrition. The analyses included mental category X attrition tables, discriminant function analysis, and two-factor multivariate analysis of variance.

There seems to be no doubt that in the case of this sample of USCG recruits, IRT scoring was no better than number-right scoring for prediction of boot camp attrition. However, it should be quickly noted that these results definitely are <u>not</u> an indictment of IRT scoring. First of all, a variable such as boot camp graduation/attrition is a very crude criterion for prediction from verbal, arithmetic, and mechanical comprehension ability levels. Also, as mentioned previously, the attrition rate was only 8.7%. There just isn't much room for improving upon a base rate prediction accuracy of 91.3% (i.e., simply predict everyone will graduate).

Another major factor may be the method used to generate the IRT scores. It is known that Bayesian modal scoring is biased toward the mean just as number-right scoring is. Consequently, the resultant IRT scores tend to have very high correlations with number-right scores. Still another factor may be the test information curves. Unfortunately, CGST test information curves are not yet available. However, some analyses done on preliminary versions of the CGST indicate that the test information curves may be flatter than intended. Tests with relatively flat information curves also tend to result in IRT scores with high correlations with number-right scores. Indeed, as noted earlier in this study, the correlations between the IRT scores and the number-right scores were extremely high - - .9897, .9786, and .9735.

To repeat our assertion at the beginning, further research of this kind is needed in many different settings. There just simply is no substitute for real-world studies in which predictor scores and criterion scores are provided by real people. This is the only way to determine whether IRT scoring will provide the true payoff for our organization - - significant improvement in predictive validity as compared to number-right scoring.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In Frederic M. Lord and Melvin R. Novick, Statistical theories of mental test scores. Reading Mass: Addison-Wesley, 1968, pp 397-424.

Hambleton, R. K. and Cook, L. L. Latent trait models and their use in an analysis of educational test data. Journal of Educational Measurement, 1977, 14: 75-96

Urry, V. W. Ancillary estimators for the item parameters of mental test models. In W. A. Gorham (Chair), Computers and testing: Steps toward the inevitable conquest. Symposium presented at the 83rd Annual Convention of the American Psychological Association, Chicago, 1975. (NTIS No. PB 261 694/AS)

Urry, V. W. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement. 1977, 14, 181-196.

Weiss, D. and Davison, M. 1981 - Review of test theory and methods. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program (Research Report 81-1)

SOLDIER/MACHINE INTERFACE EVALUATION OF ARMY DEVELOPMENTAL SYSTEMS

James C. Geddie
Chief, AMSAA Detachment
U.S. Army Human Engineering Laboratory

With operations and maintenance costs absorbing more than 80% of the total life cycle cost of a typical Army system, there is increased emphasis on an effective soldier/machine interface. The notion of "soldier/machine system" is contrasted with that of "hardware system" along with implications for testing. The three categories of data -- human engineering measurements, user acceptance/opinion data, and human performance data -- are discussed with a description of collection methods and the role of each in a soldier/machine system evaluation.

For systems recently fielded by the Army and for those still in development, it is fairly common for the life cycle cost projections for a typical system to show only about 20% of the total cost for the system to be research, development and acquisition costs. The remaining 80% is absorbed by operation, maintenance and other support costs. This latter 80% is heavily weighted with "people costs" to include trainers, operators, maintainers and others. Thus it is becoming increasingly apparent to Army planners that greater emphasis on developing systems which have an effective and efficient soldier/machine interface -- even at the expense of increasing the up-front acquisition costs -- promises soldier/machine systems whose total life cycle cost effectiveness is dramatically enhanced.

Soldier/Machine Systems vs Hardware

This increased emphasis on the soldier/machine interface has not been a sudden change. Instead it has been a gradual one coincident with and related to an increased willingness to think in terms of developing soldier/machine systems rather than developing separate hardware, training, software, logistic support, technical documentation and facilities as separate efforts to be later combined into an effective system. It should be added at this point that the metamorphosis is by no means complete. There are many members of the developmental community who still say "system" when what they really mean is "hardware".

At first encounter the difference between the two concepts may seem somewhat superficial, but it has fundamental implications for the way in which the Army develops new combat systems. It also has some profound consequences for testing of developmental systems.

During development, the systems approach requires that many players get into the process right at the beginning. The materiel developer, the combat developer, the trainer, the logistician and several others must participate in defining the goals, requirements and limitations of the system. Unlike the somewhat fragmented approach to development mentioned above, the systems approach requires that each player participate throughout the process in a kind of trade-off process. In this process an attempt is made to arrive at a cost-effective means of acquiring a new combat capability. Among the trade-offs which might, for example, be negotiated is a choice between (1) a hardware design which is high in acquisition costs but imposes human performance and skill requirements which are few and cheap to acquire and maintain versus (2) a hardware design which is lower in acquisition costs, but requires human performance and skills which may be very costly to acquire and to maintain. Based upon the available cost predictions, the alternative which offered the lowest life cycle cost could be selected. With only this very brief and general treatment of the notion of soldier/machine systems in development, we will proceed to a discussion of the consequences of this approach for testing.

Implications for Testing

Once the focus of attention shifts from "materiel" to "system", the business of testing and evaluation becomes quite different from straight materiel testing. In component level and even subsystem level materiel testing, hardware functions are exercised in a way in which any human function is assumed to have a probability of 1.0 of being correctly performed the first time and every time, regardless of the conditions surrounding the performance. The concern in this type of testing is nothing more than, "Did the hardware behave as predicted?" The assumption is explicit that failures due to human error are not chargeable to the hardware design. An implicit (and quite questionable) assumption of this approach to development and testing is that the combination of manpower and training resources available to the Army when this hardware is fielded will be capable of meeting whatever human performance requirements have been built into the system.

Testing and evaluation conducted under a systems approach acknowledges the influence of the human operator and his or her performance on total system effectiveness and reliability. Soldier-in-the-loop testing attempts to exercise the system using a sample of soldiers who by aptitude, training, experience, and physical characteristics are typical of those who will function as part of the system once it is fielded. An attempt is made to control the variation in this soldier sample by use of selection criteria and by use of demographic data (6). Within available test resources, system functions are exercised over a representative sample of the conditions (terrain, weather, visibility, etc.) which are anticipated in use of the system in training and in combat.

In the interest of efficient use of test resources, human factors testing is frequently done concurrently with testing for other aspects of system performance. For example, in testing a vehicular system, data on critical driving tasks might be collected on the same exercises used to test reliability or durability of the vehicle.

Human Factors Test Data

Whether human factors data are collected in a separate test or during testing of some other aspect of performance, there are basically three categories of data which will be collected. These are (1) human engineering measurements, (2) user acceptance/opinion data, and (3) human performance data. Each makes its own unique contribution to the evaluation and will be treated separately.

Human Engineering Measurements. When the Project Management Office or the prime contractor on a given system begins to plan for human factors testing, human engineering measurements are usually considered first. During the trade-offs mentioned earlier in this paper as taking place during establishment of the goals, requirements and limitations associated with a system, applicable sections of military standards (2,5) and specifications will have been cited as design standards to be met. The requirements document may also specify military handbooks (4) and other sources of guidance to be used in hardware design.

In general, human engineering measurements are the data used to evaluate compliance with these engineering requirements. The required data may often be collected without participation of the human component of the system and may address:

size
weight
lighting level
noise level
crew workspace layout
ingress and egress provisions
temperature
whole body vibration
brightness, ligibility and labelling of displays
placement and force requirements of controls

These are some of the physical measurements which might be made to permit evaluation against design standards and contract specifications.

User Acceptance/Opinion Data. A second category of test data collected for a human factors evaluation is used to learn from test participants, test control personnel and observers, characteristics of the system which would not be revealed by human engineering measurements. While it is generally accepted that troops will not function effectively with equipment they dislike or mistrust, a more important purpose for attempting to tap the user's experience is that the user can sometimes suggest improvements in hardware design or in operating procedures which would never be revealed by checking for compliance with specifications or standards. After an attempt to function with the hardware and procedures as a total system, the test participants' insights can be very predictive of the level of acceptance to be expected from users once the system is fielded. Their insights may also identify protential problem areas which merit close scrutiny during human performance testing.

Generally, user opinion data are collected by means of questionnaires, structured interviews and voluntary reports.

Questionnaires may be administered periodically throughout the test, with an interview conducted perhaps at or near the end of testing. Volunteered reports are recorded as they occur.

Human Performance Data. The overall goal of a human factors engineering program is to ensure the compatibility of (1) the soldier, (2) the training, (3) the tasks, and (4) the equipment. Having compared the hardware characteristics against human engineering standards and trained a sample of soldiers and elicited opinions of the equipment and training, the remaining task in determining whether the human factors engineering program has met its goal is exercising mission-critical soldier tasks and collecting and analyzing performance data on those tasks.

Collection of this category of data usually begins with a review of the human performance requirements associated with the system. If the system contract follows MIL-H-46855B and has specified contract data item DI-H-7055, Critical Task Analysis Report (3,1) as a deliverable data item, this data item is an ideal place to begin identifying tasks to exercise in collecting human performance data. Other input to the task selection process would include training materials used in training test participants, and draft technical and training manuals. The list of tasks to be excercised should have as its highest priority those tasks whose performance represents an outer limit on total system performance. An example of such a task might be laying the gun in a tank system.

There are two basic measures used in human performance testing. One is performance time; the other is error rate. For each task exercised and measured in the test, both kinds of data should be collected on each event. The reason for insisting that both measures be made of each event is that for most tasks, performance time and error rate can be traded off one for the other. The set established in the test participant can radically affect whether he or she emphasizes speed of performance at the expense of accuracy or conversely, accuracy at the expense of speed. For some tasks, the trade-off function itself might be critically important in affecting operational doctrine for the system.

Analysis of these data should first compare achieved performance against the performance goals established in the system requirements documents. If the trade-off process mentioned earlier in establishing system goals, requirements, and limitations has included a thorough treatment of human per.ormance requirements, human performance time and error rate criteria in the requirements document will serve as test criteria for performance of mission critical tasks. Where no such criteria are established, the performance data will be used in predicting system performance. The question of "How good is good enough?" then gets a *post hoc* answer, but the answer may be a more reasonable one with knowledge of currently achieved performance.

Another use -- perhaps a more important one -- for these data is in identifying areas in which human factors engineering design improvements have a high payoff potential. If, for example, the data show an unexpectedly long performance time for one of a series of sequentially performed tasks, that task would be identified as a priority candidate for improvement in the hardware associated with the task or the procedures for performing it. Consideration might also be given to automating part or all of its performance.

Soldier/Machine System Evaluation

The botton line for evaluation of a soldier/machine system has been reached when the evaluator has answered the question, "So what?", for each test issue. The specifications and standards against which we compare engineering measurements should be met. Their criteria have been developed from experience with many systems, and meeting those criteria improves the probability of acquiring an effective and efficient soldier/machine system. However, meeting those requirements does not insure that this has been achieved, nor does failure to meet one or more of the criteria insure an unsuccessful system.

Likewise, the user's evaluation is important. The Army has demonstrated too many times that a system whose design the soldiers dislike or distrust has little chance of operational success. Unfortunately (or maybe it isn't so unfortunate, except for testing), there is not a one for one correspondence between equipment characteristics about which users complain and equipment characteristics which can be shown to degrade total system performance. In competitive testing it has also been observed that soldiers often state a preference for an equipment design with which their performance is worse than with a less liked competitor.

The indications from engineering measurements and from user input are important in a system evaluation. They should influence the selection of tasks for human performance testing. Obviously, if an equipment characteristic clearly violates a human engineering standard or users identify a design characteristic that they feel significantly degrades performance, these characteristics merit closer examination. The most direct answer to the bottom line question, "So what?", comes from human performance testing. It is from these data that the evaluator can learn whether the identified system characteristics significantly affect total system effectiveness.

REFERENCES

1. Department of Defense. Critical task analysis report (DI-H-7055). Philadelphia, PA: Naval Publications and Forms Center, 1 June 1979.

2. Department of Defense. Human engineering design criteria for military systems equipment and facilities (MIL-STD-1472C). Philadelphia, PA: Naval Publications and Forms Center, 2 May 1981.

3. Department of Defense. Human engineering requirements for military systems, equipment and facilities (MIL-H-46855B). Philadelphia, PA: Naval Publications and Forms Center, 1 June 1979.

4. Department of Defense. Human factors engineering design for Army materiel (MIL-HDBK-759A). Phidadelphia, PA: Naval Publications and Forms Center, 30 June 1981.

5. Department of Defense. Noise limits for Army materiel (MIL-STD-1474B). Philadelphia, PA: Naval Publications and Forms Center, 18 June 1979.

6. Geddie, James C. Profiling the characteristics of the developmental test participant. (Technical Memorandum 31-76). Aberdeen Proving Ground, MD: U.S. Army Human Engineering Laboratory, October 1976.

AD P001320

Gilbert, Arthur C. F. & Tulloh, Nancy E., US Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia. (Wed. P.M.)

Academic Preparation, Aptitude Measures and Army Officer Performance

Earlier exploratory research indicated that officers with different academic backgrounds (i.e., college majors) performed differently on certain measures of duty performance. The purpose of this research was to extend the scope of this earlier research by evaluating the effects of academic preparation on the relevant performance measures when the differences in aptitude are controlled. A sample of officers were divided into five groups on the basis of the major field of study pursued as an undergraduate. Comparison was made among these five groups on several performance measures while using certain measures of aptitude as covariates. Discussion of the findings and implications of these findings will be presented.

ACADEMIC PREPARATION, APTITUDES MEASURES AND ARMY OFFICER PERFORMANCE

Arthur C. F. Gilbert, Ph.D.
Nancy E. Tulloh*

US Army Research Institute for the
Behavioral and Social Sciences[1]
Alexandria, Virginia 22333

Knowledge of the contribution that academic preparation makes to the performance of Army officers in actual duty performance could lead to assignment strategies that would enhance officer utilization and career progression. Earlier research (Gilbert, Waldkoetter & Castelnovo, 1978) explored the differences among officers in the Field Artillery branch who pursued different undergraduate college majors in Officer Basic Course and on the average Officer Efficiency Report (OER) ratings during the first year of active duty. The results of that research did not indicate any statistically significant differences among the different college major groups on the criterion variables. Subsequent research (Gilbert, 1980) yielded results that indicated differences among officers with different academic branches for a sample of officers from all of the 13 Career Branches on certain aptitude and duty performance measures. Within the different types of branches (i.e., Combat Arms, Combat Support and Combat Service Support branches) differences on certain measures of aptitude and duty peformance were also obtained. In both of these investigations, the analysis of variance technique was employed and consequently the effect of aptitude measures on duty performance was not controlled.

This research extends the scope of these earlier research efforts. This investigation was designed to evaluate the differences in duty performance measures while taking into acount the effect of certain aptitude measures predictive of performance on these measures. These differences were evaluated (1) within the total sample of officers in all of the 13 career branches, disregarding the type of career branch to which the officers were assigned, and (2) within groupings of the career branches (i.e., Combat Arms, Combat Support, and Combat Service Support).

METHOD

A sample of officers who completed Officer Basic Course (OBC) and who entered on a tour of active duty after completion of that course were used as subjects for this research. Officers were divided into five groups on the basis of the curriculum that they pursued as undergraduates. These college major groups were Humanities, Business, Engineering, Physical Science, and Social Science.

---

*Now with Essex Corporation, Alexandria, Virginia

[1]The views expressed in this paper are those of the authors and do not necessarily reflect the view of the US Army Research Institute or the Department of the Army.

The criterion measures used consisted of three types. The first set of criterion measures consisted of measures obtained in Officer Basic Courses (OBC); these measures were the final course grades obtained for the course and the peer ratings obtained at the end of the course. The second set of criterion measures were the ratings obtained on a specially constructed Performance Evaluation Form (Gilbert & Grafton, 1976) which was based on research reported by Helme, Willemin, and Grafton (1971), Fleishman (1974), Stogdill (1974) and Willemin (1965). This instrument yielded a global measure of duty performance and measures along nine separate dimensions of Army officer performance. Ratings on each of these dimensions were obtained after 12 to 18 months of duty performance from each officer's immediate supervisor, from a senior officer who had knowledge of the officer's performance (but not necessarily the endorsing officer for Officer Efficiency Report purposes) and from two close associates. These four ratings were averaged for each of the 10 measures obtained from the instrument.

The third set of measures consisted of the annual average Officer Efficiency Report (OER) ratings for each of the first three years of active duty and the average Officer Efficiency report rating across all three years.

The aptitude measures used as control variables were obtained from the Officer Evaluation Battery (OEB). In an earlier research effort the utility of the Officer Evaluation Battery is presented (Gilbert, 1977). The three aptitude scales of this instrument are the Combat Leadership (Cognitive), Technical Managerial (Cognitive), and Career Potential (Cognitive) scales. The multiple correlations between these three measures and each of the performance measures are shown in Table 1.

Two sets of separate but parallel analyses were performed for each of the performance measures. One set of analyses consisted of an analysis of variance for each measure. The other set of analyses consisted of an analysis of covariance for each measure using the three Officer Evaluation Battery aptitude measures as predictors. Only those officers for whom all three covariates and the relative performance measure were available were used as subjects in these analyses. These two sets of analyses were performed for the total sample disregarding career branch. Officers were then divided as basis of membership in the three types of career branches: Combat Arms, Combat Support, and Combat Service Support. The performance of the five groups of majors within each type of career branch was then compared using the analysis of variance and the analysis of covariance techniques.

RESULTS AND DISCUSSION

In Table 2 are shown the means of the five college major groups for the total sample on all performance measures and the corresponding F-ratios indicating the difference among these group means by analysis of variance. Also, shown in Table 2 are the adjusted group means for the different performance measures resulting from the analysis of covariance using the three aptitude measures as covariates and the corresponding F-ratios for these analyses.

Table 1

Multiple Correlations between The Three Aptitude Measures and Each Performance Measure

Performance Measure	Multiple Correlation
Officer Basic Course Grade	.40**
Peer Ratings	.15**
Duty Performance	.12**
Combat Leadership	.22**
Technical Managerial Leadership	.09**
Tactical Knowledge	.23**
Understanding Mission	.10**
Making Decisions	.12**
Defining Subordinate Roles	.06
Planning and Organizing	.07*
Motivating Troops	.11**
Logistical Knowledge	.13**
Annual OER Rating - 1974	.06*
Annual OER Rating - 1975	.07*
Annual OER Rating - 1976	.05
Weighted Average OER	.06*

*Significant at the .05 level.
**Significant at the .01 level.

Table 2

Means and Adjusted Means for the Five Groups of College Majors on Duty Performance Measures for the Total Sample

Variable	Group Mean					F	Adjusted Group Mean					F
	H N=292	B N=725	E N=386	PS N=1370	SS N=1428		H N=292	B N=725	E N=386	PS N=1370	SS N=1428	
Officer Basic Course Grade	99.06	99.51	106.93	100.54	97.99	11.86**	101.80	102.08	104.16	97.94	100.03	8.13**
Peer Rating Final Score	97.24	100.84	102.39	100.24	99.37	3.38**	98.07	100.78	102.27	99.86	99.63	2.21
Duty Performance	99.11	101.87	99.95	100.85	98.37	2.13	100.01	102.01	99.69	100.49	98.65	1.62
Combat Leadership	93.46	99.07	98.68	104.48	97.48	16.64**	94.96	99.03	98.66	104.05	97.75	11.66**
Technical/Managerial Leadership	98.67	102.97	102.27	100.37	97.88	4.20**	99.50	103.62	101.54	99.71	98.44	3.77**
Tactical Knowledge	92.96	97.94	98.53	104.72	97.47	19.14**	94.71	98.26	98.29	103.99	97.96	11.84**
Understanding Mission	98.42	102.15	99.62	100.87	98.33	2.53*	99.12	102.05	99.61	100.73	98.43	2.07
Making Decisions	96.68	101.40	98.71	102.13	98.13	5.06**	97.41	101.14	98.78	102.07	98.16	4.14**
Defining Subordinate Roles	100.16	101.86	98.92	100.62	98.63	1.74	100.58	101.80	98.82	100.54	98.71	1.62
Planning & Organizing	99.34	102.43	101.11	100.13	98.84	1.83	99.96	102.51	100.99	99.90	99.00	1.67
Motivating Troops	98.13	101.91	97.88	100.32	98.98	1.88	98.21	101.26	99.08	100.69	98.41	1.68
Logistical Knowledge	94.81	102.89	102.55	100.68	98.61	5.47**	95.97	103.69	101.43	99.83	99.35	4.33**
Annual OER Scores												
1974	99.74	100.52	100.08	102.00	98.42	5.61**	99.97	100.47	100.33	101.92	98.41	4.90**
1975	98.08	100.16	98.68	101.56	99.78	3.01*	98.31	99.99	99.29	101.57	99.63	2.50*
1976	100.99	99.88	97.01	101.21	98.71	3.45**	101.10	99.52	97.42	101.42	98.53	3.66**
Weighted OER Scores	100.11	100.62	100.04	101.17	98.71	2.92*	100.43	100.50	100.26	101.11	98.71	2.62*

*Indicates a significant difference among groups at the .05 level.
**Indicates a significant difference among groups at the .01 level

Sample size varies in analyses.

H- Humanities
B- Business
E- Engineering
PS- Physical Science
SS- Social Science

Differences were obtained among the five groups of college majors for the Officer Basic Course final grades using both the analysis of variance and the analysis of covariance techniques. In both analyses the differences were significant at the .01 level and the Engineering majors were favored. Significant differences were obtained among the means of the five groups for the peer ratings received in Officer Basic Course using the analysis of variance technique; these differences were significant at the .01 level and the Engineering majors were favored on this variable. However, the results of the analysis of covariance using this measure as the criterion failed to indicate any differences among the group means.

Results of the analysis of variance and the analysis of covariance yielded significant differences among the group means for five of the dimensions of duty performance as measured by the Performance Evaluation Form at the .01 level. These differences were obtained for Combat Leadership, Technical Managerial Leadership, Tactical Knowledge, Making Decisions, and Logistical Knowledge. Physical Science majors were favored in the Combat Leadership, Tactical Knowledge, and Making Decisions dimensions. However, business majors were favored on Technical Managerial and Logistical Knowledge ratings. Use of the analysis of variance technique resulted in a difference among the group means on the Understanding of Mission dimension at the .05 level on which Business majors were favored but analysis of covariance failed to show a difference among the five group means on this dimension.

On Officer Efficiency Report ratings differences were obtained among the five groups for each year average using both the analysis of variance and analysis of covariance techniques. The differences among group means were significant at the .01 level for the 1974 and 1976 annual average ratings and the .05 for the 1975 annual average rating. Differences among the group means for the three-year Officer Efficiency Report annual average at the .05 level. The mean of the Physical Science majors was highest on all four Officer Efficiency Report indices.

The means and the adjusted means resulting from the analysis of covariance for the five college major groups in the Combat Arms branches are shown in Table 3. The analysis of variance did not reveal significance among the means of the groups for Officer Basic Course final grades but the results of the analysis of covariance yielded significance at the .01 level. The mean of Engineering majors was slightly higher than that of Business majors and these two groups were favored over the others on this variable. The use of either analytic approach failed to reveal differences for the five college major groups in the Combat Arms for any of the dimensions of the Performance Evaluation Form with the exception of Logistical Knowledge. On this variable the analysis of variance approach failed to yield significance but the analysis of covariance yielded significant differences among the group means at the .05 level; the mean performance of Business majors was favored. On Officer Efficiency Report measures similar differences were obtained among the five groups using the analysis of variance and the analysis of covariance approaches in the Combat Arms branches. For the 1974 and 1975 Annual Officer Efficiency Report averages and for the three year average differences were obtained at the .01 level

Table 3

Means and Adjusted Means for the Five Groups of College Majors on Duty Performance Measures for the Combat Arms Branches

Variable	Group Mean					F	Adjusted Group Mean					F
	H N=113	B N=380	E N=110	PS N=979	SS N=759		H N=113	B N=380	E N=110	PS N=979	SS N=759	
Officer Basic Course Grade	100.57	102.32	106.37	99.84	101.55	2.12	102.60	104.28	104.75	98.30	102.64	6.23**
Peer Rating Final Score	99.04	99.68	100.39	99.36	99.16	.14	99.28	99.55	101.04	99.43	99.01	.87
Duty Performance	97.86*	103.64	102.33	100.87	100.21	1.21	98.46	103.84	102.11	100.67	100.41	1.15
Combat Leadership	100.88	105.63	101.82	105.99	103.13	2.14	101.36	105.63	101.81	105.98	103.15	1.83
Technical/Managerial Leadership	95.72	103.10	99.91	99.46	98.48	2.06	96.78	103.89	99.33	98.82	99.15	2.35
Tactical Knowledge	101.56	104.50	103.34	106.59	103.95	1.85	102.49	104.88	102.86	106.22	104.34	.98
Understanding Mission	97.50	104.49	102.62	101.13	99.77	2.02	97.96	104.68	102.43	100.95	99.96	1.93
Making Decisions	96.65	105.20	103.27	102.27	100.69	2.45*	96.89	105.17	103.38	102.28	100.64	2.37
Defining Subordinate Roles	100.21	102.83	102.29	100.44	98.92	1.17	100.44	102.92	102.04	100.34	99.06	1.09
Planning & Organizing	96.88	103.13	101.46	99.62	99.49	1.47	97.60	103.54	101.15	99.27	99.85	.17
Motivating Troops	99.97	104.46	100.91	100.52	100.53	1.31	100.13	104.46	101.53	100.58	100.33	1.31
Logistical Knowledge	97.38	103.68	99.18	99.62	99.21	1.73	98.64	104.56	98.17	98.86	100.08	2.46*
Annual OER Scores												
1974	99.46	101.97	101.92	102.90	98.90	4.13**	99.56	102.14	102.46	102.80	98.86	3.84**
1975	94.74	101.42	100.51	102.50	99.88	4.60**	94.90	101.44	101.21	102.49	99.77	4.31**
1976	101.67	100.93	98.43	101.86	98.52	2.40*	101.69	100.73	98.90	101.98	98.37	2.47*
Weighted OER Scores	98.80	101.80	99.89	102.00	98.56	4.10**	98.90	101.77	100.40	102.03	98.45	4.09**

*Indicates a significant difference among groups at the .05 level.
**Indicates a significant difference among groups at the .01 level

Sample size varies in analyses.

H- Humanities
B- Business
E- Engineering
PS- Physical Science
SS- Social Science

while for the 1976 annual average the differences among group means was significant only at the .05 level. The means of the Physical Science majors was favored on all of these indices.

The results of the analysis of variance and for the analysis of covariance is shown in Table 4 for the five college groups in the Combat Support branches. Results of both the analysis of variance and of analysis of covariance yielded signficant differences on Officer Basic Course grades at the .01 level, and in both analyses, the mean of engineering majors was favored. On the Combat Leadership and Tactical Knowledge dimensions of the Performance Evaluation Form, differences were obtained among the means of the five groups at the .01 level, and in both instances, physical science majors were favored. Significance was also obtained at the .01 level in the analysis of variance on the Technical-Managerial Leadership dimension, and at the .05 level in the analysis of covariance on the Decision-Making dimension. Engineering majors were favored on the Technical-Managerial Leadership dimension, and physical science majors were favored on the Decision-Making dimension. Differences were not obtained among the means of the five college majors on the four Officer Efficiency Report indices.

The results of the analyses for the five college major groups are shown in Table 5 for the officers in the Combat Service Support branches. The difference among the group means on Officer Basic Course final grades was significant at the .01 level using the analysis of variance approach; the mean of the Engineering group was highest. However, the results of the analysis of covariance indicated differences significant at the .05 level and the adjusted mean of thie Business majors was highest. Results of the analysis of variance indicated significance among the groups for peer ratings at the .05 level on which Physical Science majors were favored; however, the results of the analysis of covariance failed to indicate differences among the groups on that measure. Analysis of variance results yielded differences at .05 level for three of the dimensions of the Performance Evaluation Form. These differences were obtained for the Combat Leadership, Tactical Knowledge, and Logistical Knowledge dimensions and Physical Science majors were favored on all three dimensions. The results of the analysis of covariance, however, failed to indicate any difference among the five groups in the Combat Service Support branches on those measures. Neither the results of the analysis of variance or of the analysis of covariance indicated any differences among the means of the five groups on the Officer Efficiency Report annual ratings or for the three-year average of Officer Efficiency Report ratings.

In summary, Engineering majors received higher Officer Basic Course grades in the total sample, in the Combat Arms branches and in the Combat Support branches while Business majors received higher Officer Basic course grades in the Combat Service Support branches. This may be due to the relevance of these college curricula to the curricula of the Officer Basic Courses. Business majors were favored on the Technical Managerial ratings in the total sample and in the logistical knowledge ratings in the total sample and in the Combat Arms branches. These results are to be expected since both dimensions are characteristic of staff-managerial functions. However, Physical Science majors were favored on logistical knowledge ratings in the Combat Service Supprt branches. Physical Science majors

Table 4

Means and Adjusted Means for the Five Groups of College Majors on Duty Performance Measures for the Combat Support Branches

Variable	Group Mean					F	Adjusted Group Mean					F
	H N-97	B N-110	E N-229	PS N-271	SS N-372		H N-97	B N-110	E N-229	PS N-271	SS N-372	
Officer Basic Course Grade	97.71	94.52	107.71	102.08	95.39	14.72**	101.34	97.04	104.46	98.70	98.63	4.34**
Peer Rating Final Score	98.35	103.91	103.96	102.26	100.76	1.53	99.49	103.66	103.39	101.72	101.34	.71
Duty Performance	100.65	100.41	99.50	100.58	95.62	1.41	101.21	99.64	99.18	100.78	95.75	1.25
Combat Leadership	94.13	99.14	100.43	102.81	92.99	5.72**	94.85	97.85	100.25	103.21	92.99	5.18**
Technical/Managerial Leadership	99.64	100.59	104.15	102.37	95.17	3.53**	100.54	100.66	103.01	101.92	96.09	1.88
Tactical Knowledge	94.17	100.14	99.84	102.04	92.60	5.59**	94.80	99.27	99.73	102.16	92.70	4.92**
Understanding Mission	100.70	103.40	100.22	100.05	95.75	1.71	101.09	102.33	100.31	100.50	95.55	1.62
Making Decisions	98.83	99.53	97.87	102.16	95.49	2.15	99.21	98.13	97.90	102.91	95.15	2.63*
Defining Subordinate Roles	98.82	100.45	99.12	101.03	96.84	.84	99.23	99.53	98.98	101.44	96.76	.87
Planning & Organizing	99.40	101.13	101.56	101.87	97.38	1.04	99.47	99.94	102.10	102.55	96.79	1.39
Motivating Troops	98.96	102.39	98.94	100.34	96.27	1.17	98.80	100.80	99.92	101.37	95.24	1.70
Logistical Knowledge	93.99	102.28	103.87	102.95	97.41	3.13	94.60	102.04	103.22	102.81	97.90	2.10
Annual OER Scores												
1974	101.78	99.28	100.42	100.58	97.81	1.35	102.21	98.71	100.12	100.84	97.87	1.42
1975	101.52	100.24	98.33	100.51	101.14	1.03	101.81	100.13	98.39	100.41	101.12	.90
1976	99.41	97.63	95.34	99.29	100.20	1.94	99.44	97.07	95.22	99.92	99.81	1.96
Weighted OER Scores	102.50	98.66	100.65	100.82	99.69	.66	103.02	98.23	100.25	100.89	98.89	.87

*Indicates a significant difference among groups at the .05 level.
**Indicates a significant difference among groups at the .01 level

H= Humanities
B= Business
E= Engineering

Table 5

Means and Adjusted Means for the Five Groups of College Majors on Duty Performance Measures for the Combat Service Support Branches

Variable	Group Mean					F	Adjusted Group Mean					F
	H N=82	B N=235	E N=47	PS N=120	SS N=297		H N=82	B N=235	E N=47	PS N=120	SS N=297	
Officer Basic Course Grade	98.71	96.62	103.92	101.79	90.27	5.81**	98.96	100.67	97.90	96.75	92.66	2.63*
Peer Rating Final Score	93.21	101.42	98.99	103.06	98.25	2.97*	94.03	101.33	97.17	101.26	99.08	1.85
Duty Performance	99.26	100.32	97.74	101.34	96.91	.66	99.95	100.31	97.04	100.51	97.26	.48
Combat Leadership	83.39	90.51	86.79	93.76	87.29	2.56*	84.96	90.13	85.80	92.37	88.06	1.34
Technical/Managerial Leadership	101.53	104.00	98.98	104.07	99.99	.80	102.45	103.81	98.48	103.08	100.48	.55
Tactical Knowledge	80.88	88.30	85.54	93.23	85.40	3.27*	82.40	87.85	84.75	91.82	86.19	1.75
Understanding Mission	97.48	98.50	92.35	100.58	97.78	.79	98.87	97.64	92.47	99.93	98.44	.67
Making Decisions	94.72	97.44	94.40	100.59	94.40	1.11	96.16	96.98	93.60	99.42	95.11	.56
Defining Subordinate Roles	101.34	101.32	92.55	101.25	100.28	1.19	101.97	101.02	92.49	100.80	100.58	1.09
Planning & Organizing	102.42	102.20	98.81	100.50	98.99	.51	103.68	101.99	97.76	99.24	99.62	.60
Motivating Troops	95.01	98.40	88.80	98.24	98.30	1.54	95.32	96.67	92.27	99.80	98.32	.88
Logistical Knowledge	92.30	102.17	103.06	105.11	98.52	2.97*	93.84	102.68	100.41	102.59	99.34	1.55
Annual OER Scores												
1974	97.67	98.91	94.20	97.99	98.00	.74	97.81	98.81	94.19	97.95	98.06	.65
1975	98.46	97.93	95.54	95.41	97.79	.35	98.77	97.26	97.53	96.66	97.48	.09
1976	101.62	98.76	103.62	97.80	97.47	.97	101.78	98.48	104.53	98.21	97.41	1.12
Weighted OER Scores	99.10	99.62	97.45	95.16	97.88	.98	99.35	99.40	97.61	95.13	97.97	.85

*Indicates a significant difference among groups at the .05 level.
**Indicates a significant difference among groups at the .01 level.

Sample size varies in analyses.

H= Humanities
B= Business
E= Engineering
PS= Physical Science

received higher tactical knowledge ratings in the total sample, in the Combat Support and in the Combat Service Support branches and were rated higher on making decisions in the total sample and in the Combat Support branches; they also had a higher group mean on the Officer Efficiency Report (OER) indices in the total sample and in the Combat Arms branches.

The results of this research tend to indicate that certain college curricula may be more compatable than others with officer duty performance within the different types of Army career branches. The fact that these findings are not clear-cut may be due to the fact that the grouping of branches used did not provide for a homogeneous set of performance requirements and that the grouping of college majors were not sufficiently refined. Future research in this area will take those considerations into account.

REFERENCES

Fleishman, E. A. Twenty years of consideration and structure. In Fleishman, E. A. and J. G. Hunt (Eds.) Current development in the study of leadership. Carbondale, Ill.: Southern Illinois University Press, 1974.

Gilbert, A. C. F. Predictive utility of the Officer Evaluation Battery (OEB). Paper presented at the 20th Annual Conference of the Military Testing Association, Oklahoma City, OK, October 30-Novmber 3, 1978. In Proceedings, 20th Annual Conference of the Military Testing Association. Oklahoma City, OK: U.S. Coast Guard Institute, November 1978.

Gilbert, A. C. F. College major and Army officer performance. Paper presented at the 22nd Annual Conference of the Military Testing Association Toronto, Canada. In Proceedings of the 22nd Annual Conference of the Military Testing Association. Toronto, Canada: Canadian Forces Personnel Applied Research Unit, October 1980.

Gilbert, A. C. F. & Grafton, F. C. Some properties of an officer measure of performance and potential. (Technical Paper 345). Alexandria, VA: US Army Research Institute for the Behaviorsl and Social Sciences, October 1978.

Gilbert, A. C. F., Waldkoetter, R. O., & Castelnovo, A. E. Prediction of Field Artillery officer performance. Paper presented at the 20th Annual Conference of the Military Testing Association, Oklahoma City, OK, October 30-November 3, 1978. In Proceedings, 20th Annual Conference of the Military testing Association. Oklahoma City, OK: U.S. Coast Guard Institute, November 1978.

Helme, W. H. Willemin, L. P. & Grafton, F. C. Dimensions of leadership in a simulated combat situation. (Technical Research Report 1172). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, July 1971. (NTIS No. AD 730 315)

Stogdill, R. M. Handbook of leadership: A survey of theory and research. New York: The Free Press, 1974.

Willemin, L. P. Criterion aspects of Army research on the prediction of officer performance. (Research Study 65-6). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences, December, 1965.

AD P001321

Gilbert, Arthur C.F. & Tulloh, Nancy E., US Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia. (Wed. P.M.)

Army Officer Success in the Different Types of Career Branches

The purpose of this research was to evaluate the utility of different measures of potential in identifying groups of Army officers who were defined as being successful in different types of career branches. Several measures of aptitude and motivation were used in the analyses. Implications of the findings will be presented.

ARMY OFFICER SUCCESS IN THE DIFFERENT TYPES OF CAREER BRANCHES

Arthur C. F. Gilbert, Ph.D.
Nancy E. Tulloh*

US Army Research Institute for the
Behavioral and Social Sciences[1]
Alexandria, Virginia 22333

In previous research, the characteristics of those officers who received higher final course grades in Officer Basic Courses (OBC) than would have been predicted on the basis of aptitude measures were explored (Gilbert, 1980). The results of that research indicated that those officers displayed a greater interest in becoming an Army officer and a greater interest in those activities related to success as a combat leader as these attributes were reflected by the relevant scales of the Officer Evaluation Battery (OEB). Other research has explored the utility of certain measures in the prediction of different aspects of duty performance (Gilbert, 1977). While the research indicated the utility of certain of these measures, there were certain differences in the utility of these predictors in the different types of career branches (i.e., Combat Arms, Combat Support, and Combat Service Support). These research efforts then lead to the question of whether success in the different types of assignments is related to different patterns of aptitude, motivation, and ratings on early performance on active duty. If so, this information could be used for assignment purposes to insure that officers are assigned to those duty positions in which their aptitude and motivation are most fully utilized. This investigation was designed to explore these considerations.

Specifically, the purpose of this research was to explore differences that might exist among groups of officers who received higher than average ratings in the different types of career branches. These differences were evaluated on certain measures of aptitude and motivation as well as performance along certain leadership dimensions early in their tour of active duty. The utility of the variables in differentiating among these groups of officers was evaluated individually and in certain combinations.

METHOD

A sample of officers who attended Officer Basic Courses (OBC) in thirteen Army Career Branches and who continued on active duty after completion of Officer Basic Course were used as subjects in the research. A weighted average of the Officer Efficiency Reports received by these officers for the first three years of active duty was computed and an

---

*Now at Essex Corporation, Alexandria, Virginia.

[1]The views expressed in this paper are those of the authors and do not necessarily reflect the view of the U.S. Army Research Institute or the Department of the Army.

average of these weighted averages was obtained. Officers who received a greater than average weighted average were selected as subjects. This group was then divided on the basis of membership in the three types of Career Branches: Combat Arms, Combat Support, and Combat Service Support. The three groups of officers were then compared on two different types of measures.

The first set of measures consisted of measures obtained in Officer Basic Course, and included the Officer Evaluation Battery, final course grades in Officer Basic Course, and peer ratings received at the end of the Officer Basic Course. The Officer Evaluation Battery (OEB) reflects measures of aptitude and motivation in seven scale scores. The predictive utility of this instrument is discussed in an earlier paper (Gilbert, 1978) and a brief description of the scales is presented in Figure 1.

The second set of measures consisted of duty performance measures obtained from a specially constructed Performance Evaluation Form (Gilbert and Grafton, 1976) which yields a global measure of duty performance and measures along nine dimensions of duty performance. Ratings on the Performance Evaluation Form were obtained after twelve to eighteen months of duty performance.

The first series of analysis consisted of comparing the means of the three groups of officers (i.e., Combat Arms, Combat Support, and Combat Service Support) on each of the available measures. The discriminant analysis technique was used to evaluate the efficacy of all of these variables in defining group membership. Two other discriminant analyses were performed; one of these analyses used only the Officer Basic Course measures while the other used only the duty performance measures.

The second series of analyses was performed as an exploratory strategy to evaluate what the effect on group membership prediction would be if only two groups of officers were involved (i.e., differentiation between Combat Arms officers and officers not in the Combat Arms). For this series of analyses, the subjects in the Combat support and in the Combat Service Support were combined. Thus, for each of the individual measures, differences between the mean of officers in the Combat Arms branches and the mean of officers in the other branches were evaluated and a series of discriminant analyses paralleling the three-group design were performed.

RESULTS AND DISCUSSION

The means of the three groups of officers in the three types of Career Branches are shown in Table 1 for the measures of aptitude and motivation of the Officer Evaluation Form Battery (OEB). Significant differences were found among the three groups on three scales of the OEB at the .01 level; these differences were obtained for the Combat Leadership (Cognitive), Technical Managerial (Non-cognitive), and the Career Potential (Non-cognitive) scales. A statistical difference was also obtained among

SUBTEST	DESCRIPTION OF ITEMS
Combat Leadership (Cognitive)	Military tactics; practical skills in a variety of areas ranging from out-door activities to mechanical and electronic applications.
Technical-Managerial Leadership (Cognitive)	History, politics; culture; mathematics; physical sciences
Career Potential (Cognitive)	Technological knowledge relevant to military requirements.
Combat Leadership (Non-Cognitive)	Combat leader qualities, occupational interests, sports interest, outdoor interests related to combat leadership
Technical-Managerial Leadership (Non-Cognitive)	Mathematics and physical sciences skills and interest; urban or rural background; scientific interest and ability; decisive leader qualities; and verbal-social leadership
Career Potential	Clerical-administrative interest, versus white collar interest, combat interest
Career Intent	Intention of making the Army a career choice

Figure 1. Officer Evaluation Battery (OEB) Subtests and Description of Items.

Table 1

Means for the Three Groups of Officers on Officer Basic Course (OBC) Measures

Variable	Combat Arms (N=439)	Combat Support (N=139)	Combat Service Support (N=68)	Total (N=637)	F-Ratio
Combat Leadership (Cognitive)	110.84	104.97	97.37	108.12	16.13**
Technical Manage-ial Leadership (Cognitive)	111.21	110.27	102.94	110.12	4.57*
Career Potential (Cognitive)	104.09	105.46	101.94	104.16	.71
Combat Leadership (Non-Cognitive)	111.90	109.94	106.09	110.85	3.13*
Technical Managerial (Leadership (Non-Cognitive)	103.44	110.43	107.07	105.35	6.46**
Career Potential (Non-cognitive)	110.96	105.65	94.07	108.00	28.13**
Career Intent	115.99	117.48	112.85	115.98	1.85
Officer Basic Course Final Grades	104.34	103.16	100.75	103.70	1.15
Final Course Peer Ratings	99.14	105.87	109.14	101.67	13.93**

* Significant at the .05 level
** Significant at the .01 level

the three groups on the Combat Leadership (Non-cognitive) and on the Technical-Managerial (Non-cognitive) scales at the .05 level. The mean scores for officers in the Combat Arms branches were highest on the Combat Leadership (Cognitive) scale, Technical-Managerial (Cognitive) scale Combat Leadership (Non-Cognitive) scale and on the Career Potential (Non-Cognitive) scale. The mean performance of that group was lowest on the Technical-Managerial (Non-Cognitive) scales.

Statistically significant differences were not obtained among the three groups of officers on the final course grades obtained in Officer Basic Course. The differences among the three groups means were statistically significant at the .01 level on peer ratings obtained in Officer Basic Course; officers in the Combat Service Support branches had the highest mean on this variable.

The ratings of the three groups of officers on the dimensions of the Performance Evaluation Form are shown in Table 2. Statistically significant differences were obtained among the group means on the Combat-Leadership scale and the Tactical Knowledge scales at the .01 level and on the Technical-Managerial Leadership ratings at the .05 level. The mean ratings of officers in the Combat Arms group was highest for Combat Leadership and for the Tactical Knowledge ratings. Officers in the Combat Service Support branches were favored in the mean ratings on the Technical-Managerial scale. The discriminant analysis technique used to test the efficacy of the variables in predicting group membership yielded the classification data shown in Table 3. The first analysis involving all variables (i.e. Officer Basic Course measures and measures of on-the-job performance yielded 73.2 percent correct classification. Use of the Officer Basic Course variables alone yielded correct classification of 71.6 percent. The duty performance measures obtained from the Performance Evaluation Form also yielded an index of correct classification of 69.7 percent.

In the second series of analyses officers in the Combat Support branches and the Combat Sevice Support branches were combined into one group and compared with the officers in the Combat Arms groups. In Table 4, the means for the Combat Arms officers and other officers are presented for the Officer Basic Course (OBC) measures. The mean for the Combat Arms group was significantly higher at the .01 level on the Combat Leadership (Cognitive) and the Corcer Potential (Non-cognitive) scales of the Officer Evaluation Battery and at the .05 level on the Combat Leadership (Non-cognitive) scale of that instrument. Officers not assigned to the Combat Arms branches had a significantly higher (.01 level) mean on the Technical-Managerial (Non-cognitive) scale. The mean of those officers not in the Combat Arms branches was significantly higher (.01 level) for peer ratings than was the mean of those officers in the Combat Arms branches.

In Table 5, the means for the two groups of officers on duty performance measures are shown. The mean performance of officers in the Combat Arms branches is higher at the .01 level on combat leadership and tactical knowledge ratings and at the .05 on the decision making ability ratings. On the other hand officers who were not in the Combat Arms branches were favored in terms of mea- rating on the technical managerial dimension.

Table 2

Means for the Three Groups of Officers on
Measure of Duty Performance

Variable	Combat Arms (N=439)	Combat Support (N=139)	Combat Service Support (N=68)	Total (N=637)	F-Ratio
Duty Performance	108.58	108.06	107.07	108.31	.27
Combat Leadership	111.58	106.40	97.67	108.96	24.66**
Technical Managerial Leadership	104.93	106.26	111.00	105.87	3.78*
Tactical Knowledge	110.78	104.69	94.82	107.74	28.79**
Understanding Mission	107.93	106.05	105.65	107.28	1.02
Making Decisions	108.71	106.33	105.15	107.81	2.11
Defining Personal Roles	107.00	105.56	107.72	106.76	.47
Planning and Organizing	105.76	107.14	108.24	106.33	.79
Motivating Troops	106.73	105.16	103.95	106.09	.99
Logistical Knowledge	105.29	105.98	105.46	105.46	.08

*Significant at the .05 level
**Significant at the .01 level

Table 3

Three Group Classification

All Variables:

Actual	Predicted: Combat Arms	Predicted: Combat Support	Predicted: Combat Service Support
Combat Arms	94.4	3.7	1.9
Combat Support	69.8	19.4	10.8
Combat Service Support	30.9	20.6	48.5

Percent of Cases Correctly Classified = 73.2

Officer Basic Course Variables:

Actual	Predicted: Combat Arms	Predicted: Combat Support	Predicted: Combat Service Support
Combat Arms	97.9	0.7	1.4
Combat Support	79.1	12.2	8.6
Combat Service Support	61.8	11.8	26.5

Percent of Cases Correctly Classified = 71.6

Duty Performance Measures:

Actual	Predicted: Combat Arms	Predicted: Combat Support	Predicted: Combat Service Support
Combat Arms	96.3	.9	2.8
Combat Support	89.2	2.9	7.9
Combat Service Support	58.8	2.9	38.2

Percent of Cases Correctly Classified = 69.7

Table 4

Means for the Two Groups of Officers on
Officer Basic Course (OBC) Measures

Variable	Combat Arms (N=430)	Others (N=207)	Total (N=637)	F-Ratio
Combat Leadership (Cognitive)	110.84	102.47	108.12	25.17**
Technical Managerial Leadership (Cognitive)	111.21	107.85	110.12	3.54
Career Potential (Cognitive)	104.09	104.30	104.16	1.64
Combat Leadership (Non-Cognitive)	111.90	108.68	110.85	4.26*
Technical Managerial (Leadership (Non-Cognitive)	103.44	109.33	105.35	11.67**
Career Potential (Non-cognitive)	110.96	101.85	108.00	35.78**
Career Intent	115.99	115.96	115.98	.01
Officer Basic Course Final Grades	104.34	102.37	103.70	1.54
Final Course Peer Ratings	99.14	106.94	101.67	26.33**

* Significant at the .05 level
** Significant at the .01 level

Table 5

Means for the Two Groups of Officers on
Measures of Duty Performance

Variable	Combat Arms (N=430)	Others (N=207)	Total (N=637)	F-Ratio
Duty Performance	108.58	107.73	108.31	.37
Combat Leadership	111.58	103.53	108.96	34.93**
Technical Managerial Leadership	104.93	107.81	105.87	3.99*
Tactical Knowledge	110.78	101.45	107.74	41.23**
Understanding Mission	107.93	105.92	107.28	2.01
Making Decisions	108.71	105.94	107.81	3.99*
Defining Personal Roles	107.00	106.27	106.76	.24
Planning and Organizing	105.76	107.50	106.33	1..39
Motivating Troops	106.73	104.71	106.09	1.77
Logistical Knowledge	105.29	105.81	105.46	.11

*Significant at the .05 level
**Significant at the .01 level

In Table 6 the classification of the two groups of officers using the discriminant analysis technique is shown. When all of the variables were employed (i.e., Officer Basic Course measures and measures of duty performance), 80 percent correct classification resulted. The use of the Officer Basic Course measures alone resulted in 76.1 percent correct classification while use of the duty performance dimensions resulted in 73.9 percent correct classification.

The results of this research indicate that officers in the Combat Arms branches who receive higher than average Officer Efficiency Report ratings displayed a greater aptitude for success in combat type situations than did the other officers. This is reflected in their scores on the Combat Leadership (Cognitive) scale of the Officer Evaluation. However, officers in the Combat Arms branches also score higher on those aptitudes necessary for a staff or technical position as measured by the Technical-Managerial scale of the same instrument but the differences are not as pronounced. When these officers were compared to all other officers on this measure, a statistically significant difference was not obtained. The group of higher rated officers also displayed a greater interest in a career as an Army officer as measured by the Career Potential scale of the Officer Evaluation Battery and interest in those activities related to success as a combat leader as measured by the Combat Leadership (Non-Cognitive) scale of this instrument. Interest in activities related to staff and technical types of activities characterized those officers who were not in the Combat Arms branches. Officers who received higher than average Officer Efficiency Report ratings were rated higher than other officers early in their active duty tour on combat leadership and tactical knowledge while other officers were rated higher on technical-managerial leadership.

The measures used in this research were of utility in identifying officers who received higher than average Officer Efficiency Report ratings in the Combat Arms from other officers who received higher ratings in other branches. Future research will be directed toward exploring the relative merits of the indices used in this exploratory research in differentiating among officers who are highly rated in more specific officer specialties.

Table 6

Two Group Classification

All Variables:

Actual	Predicted: Combat Arms	Predicted: Others
Combat Arms	92.3	7.7
Others	48.8	51.2

Percent of Cases Correctly Classified = 79.0

Officer Basic Course Variables:

Actual	Predicted: Combat Arms	Predicted: Others
Combat Arms	93.0	7.0
Others	58.9	41.1

Percent of Cases Correctly Classified = 76.0

Duty Performance Measures:

Actual	Predicted: Combat Arms	Predicted: Others
Combat Arms	91.4	8.6
Others	62.3	37.7

Percent of Cases Correctly Classified = 73.9

REERENCES

Gilbert, A. C. F. Efficacy of certain measures in predicting Army officer performance. Paper presented at the 19th Annual Conference of the Military Testing Association, San Antonio, TX, October 17-21, 1977. In Proceedings, 19th Annual Conference of the Military Testing Association. San Antonio, TX: Air Force Human Resources Laboratory and USAF Occupational Measurement Center, October 1977.

Gilbert, A. C. F. Predictive utility of the Officer Evalution Battery (OEB). Paper presented at the 20th Annual Conference of the Military Testing Association, Oklahoma City, OK, October 30-November 3, 1978. In Proceedings, 20th Annual Conference of the Military Testing Association. Oklahoma City, OK: U.S. Coast Guard Institute, November 1978.

Gilbert, A. C. F. College major and Army officer performance. Paper presented at the 22nd Annual Conference of the Military Testing Association Toronto, Canada. In Proceedings of the 22nd Annual Conference of the Military Testing Association. Toronto, Canada: Canadian Forces Personnel Applied Research Unit, October 1980.

Gilbert, A. C. F. & Grafton, F. C. Characteristics of an officer performance evaluation measure. Paper presented at the Annual Conference of the Military Testing Association, Gulf Shores, Ala., October 18-22, 1976. In Proceedings of the 18th Annual Conference of the Military Testing Association. Pensacola, Fla: Naval Education and Training Program Development Center, 1976.

USE OF CODAP SYSTEM 80 TASK MODULES IN OCCUPATIONAL ANALYSIS

Doug Goodgame, Ph.D.
Research Specialist
Texas A&M University

The field of occupational analysis benefits from the use of CODAP which is a computerized approach for processing task inventory data. CODAP is an acronym for Comprehensive Occupational Data Analysis Programs and produces printouts which job analysts have found invaluable in understanding and interpreting work performed by individuals in a wide range of career fields. The recent development of CODAP System 80 greatly expands the job analysts' ability to process task data. The new features of System 80 overcome certain restrictions and limitations inherent in the older version of CODAP. One restriction limits the number of task modules (duty fields) in a study to 26, and once the task modules are defined and categorized, they can not be changed. CODAP System 80 permits wide flexibility in defining the modules for which tasks are to be grouped. This paper outlines the need to re-examine the process by which we define duty fields for occupational analysis. Traditional methods hinder job analysts' ability to make accurate interpretations of data and cause job analysis to spend work time on unproductive tasks.

USE OF CODAP SYSTEM 80 TASK MODULES IN OCCUPATIONAL ANALYSIS

Doug Goodgame, Ph.D.
Research Specialist
Texas A&M University

The field of occupational analysis is now two decades old and has experienced many changes, developments, and improvements in the quality of task inventory content and the computer software for processing task inventory data. Occupational analysis is the official procedure by which all branches of the US Armed Services collect and analyze job and task information to guide decision makers in developing training, testing and occupational classifications. The adoption of the task inventory approach by local, state and federal agencies, power utilities and private corporations is increasing at a rapid rate. This is because the occupational analysis process meets the requirements defined by the federal uniform guidelines for conducting job analysis prior to developing personnel selection exams. Occupational analysis is the most widely used method for developing job analytic data of all major job analysis methods.

The foundation for occupational analysis was laid in the 1960's through completion of research studies which contributed to the development of basic methodologies for collecting and analyzing task inventory data. In 1960, McCormick and Ammerman published their report entitled "Development of Work Activity Check Lists for Use in Occupational Analysis" and established the utility in partitioning jobs into task statements for collecting job information directly from incumbent workers. As a result, the US Air Force, which sponsored this study, incorporated the list of task activities into a comprehensive job survey instrument known as a job or task inventory. In time, task inventories for many occupational fields grew to contain 400 to 600 tasks thus presenting to the occupational analyst very large amounts of detailed data for analysis. Consequently, a way was needed by which summary statistics could be developed for groups of tasks known as duty fields. Mayo published a report in 1969 entitled "Three Studies of Job Inventory Procedures: Selecting Duty Categories, Interviewing, and Sampling." In this study, Mayo evaluated "alternative variations in categorization of tasks" and developed recommended duty categories for grouping task statements. The report suggested three stable formats for organizing non-supervisory tasks. These include the work section, work function, and equipment use formats. In regards to the work section format, Mayo recommended tasks be arranged in groups parallel to the work sections of organizational charts. Accounting and finance may typically be divided into travel, pay accounts, control, collecting and stock funds sections. Hence tasks could be grouped according to these sections.

Using the work function format, an inventory developer should create groupings of tasks according to major work functions such as equipment maintenance. This is most appropriate when one piece of equipment such as a helicopter is the focus of work. Typical duty fields' headings should include inspection, troubleshooting, adjusting and repairing equipment items. In this manner, tasks are grouped because of some cohesive common denominator relative to work function. In the equipment use format workers are responsible for maintenance of different types of equipment and Mayo recommends grouping tasks under duty titles such as "maintaining generators", "maintaining intake and exhaust systems" and "maintaining diesel engines". Here the work of an individual may encompass activities associated with several types of equipment.

The study did not address how the number of tasks in a duty field, or a task module as it is called in CODAP System 80, would affect the analysis of data and the work load for the job analysts. Often task inventories are developed by a group with no responsibility for analyzing data. When this occurs, problems usually result since a duty field containing a high number of tasks can present problems in analysis and interpretation for the job analysts. Typically these problems accrue because the task inventory developer failed to follow conventional guidelines and grouped tasks under broad duty fields containing tasks which in reality represent diverse work sections of an organization, work functions or equipment use. A review of these duty fields usually reveals that tasks could be further divided into smaller duty fields bringing more precision to the analysis and decreased mental strain to the analyst's work.

The following example illustrates the need for revised thinking about how we group tasks into duty fields. Figure 1 contains a duty field entitled "Performing Systems Design Functions" and reflects the "functional approach" for categorizing tasks into duty fields. The study was conducted with all twenty-six tasks in the one duty field. A close examination of the individual tasks reveals the tasks can be further subdivided into two groups. Figure 2 reveals that one group of tasks relates to the development of system procedures and specifications for a proposed system. The second group defines the types of tasks involved in working and interacting with programmers and other systems analysts in designing a new system. It is apparent that office study and research to develop the procedures and specifications for a new system requires a very different set of behaviors than managing and directing other staff members. The extent to which a systems analyst would perform tasks in either of the two groups could be determined in the analysis of data in one of two ways. If the duty field categorization had been defined and titled to reflect these two groups then the duty level job descriptions would report workers performance regarding each set of tasks. If all tasks are grouped under one category then the analyst must work through the task level job description and locate and separate the different sets of tasks. This is not an enjoyable process as any job analyst knows.

There is clearly a need for a revised look at how we classify tasks into duty categories and how this classification affects the work of job analysts and the accuracy of their analysis. In many cases, task inventory developers are unaware of the difficulties created by random methods for categorizing duties. By using CODAP System 80, we are no longer tied to using one approach for categorizing tasks and assigning them to duty fields. In fact, it is often desirable to develop different duty categories based upon planned use of data. Our work has shown that classification and training each require separate schemes for assigning tasks to duty fields. As revealed by the example, it is also desirable to construct duty fields containing tasks representing more explicit groups of behavior. This would be helpful in classification where duty level job descriptions pinpoint the nature of work speciality more precisely. Such cannot be readily determined using duty fields containing tasks representing diverse types of behavior, more than one work section, or work function. In regards to training, curricula for a training program are often divided into training modules that represent various categories of instruction. Approaches in training evaluation require that each task in an inventory be assigned to a module in which the training of the module prepares a worker to perform the task. In this situation, long training programs may encompass a large number of modules and require that tasks be apportioned accordingly.

These are but a few of the situations that demand we take a fresh look at how we define duty categories for task inventories. The work of occupational analysts in the 1980's will require more efficient and effective methods for processing and analyzing task inventory data to meet user needs.

Figure 1

DESIGNING SYSTEMS

ADVISE PROGRAMMERS AND SYSTEMS ANALYSTS ON DIVISION'S POLICY AND PROCEDURES FOR DATA PROCESSING.

ANALYZE AND DETERMINE THE TYPE OF FORMS REQUIRED TO COLLECT DATA.

ANALYZE THE LOGIC FUNCTIONS REQUIRED IN THE SPECIFICATIONS FOR WRITING PROGRAMS.

ASSIGN ANALYSTS TO PREPARE A PROPOSAL FOR PRESENTATION TO A REVIEW BOARD.

ASSIGN PERSONNEL TO PROJECT QUALITY REVIEW BOARDS (QRBS).

ASSIGN PERSONNEL TO PROJECTS.

ASSIGN SYSTEMS AND PROGRAMMING PROJECTS TO DATA PROCESSING SUPPORT TEAMS.

ASSIGN TASKS TO PROGRAMMER.

CONDUCT HARDWARE CAPACITY PLANNING REVIEWS.

DEFINE THE FUNCTIONS TO BE ASSIGNED TO PROGRAMMERS.

DETERMINE THE LEVEL OF PERSONNEL NECESSARY TO DEVELOP THE SYSTEM.

DETERMINE THE TYPES OF EQUIPMENT REQUIRED TO SUPPORT A SYSTEM.

DETERMINE THE TYPES OF PROGRAMS AND ROUTINES THAT WILL BE REQUIRED TO DEVELOP A SYSTEM.

DETERMINE TIME ESTIMATES TO COMPLETE TASKS.

DEVELOP WORK SCHEDULES FOR PROJECT.

PARTICIPATE IN WALKTHROUGHS AND REVIEWS OF PROJECTS AND/OR SYSTEMS.

PREPARE WORKPLANS FOR PROJECT.

REVIEW A PROPOSED PROJECT WITH SYSTEMS ANALYST AND USER TO DEFINE THE SCOPE OF THE PROJECT.

REVIEW AND EVALUATE DATA AND INFORMATION COLLECTION PROCEDURES.

REVIEW AND EVALUATE SPECIFICATIONS FOR A PROPOSED SYSTEM.

REVIEW FINDINGS OF FEASIBILITY STUDIES TO DETERMINE APPROPRIATE ACTIONS.

REVIEW MANUALS FOR PROPOSED SOFTWARE.

REVIEW PROJECT SCHEDULES AND ESTABLISH PRIORITIES FOR PROJECT START-UPS.

REVIEW PROPOSED STAND[illegible]S FOR SOFTWARE.

REVIEW SYSTEMS DESIGN SPECIFICATION TO ASSESS ADEQUACY OF APPROACH.

REVIEW SYSTEMS DEVELOPMENT PLANS TO DETERMINE WHAT ADDITIONAL STAFF AND RESOURCES ARE NECESSARY.

Figure 2

DESIGNING SYSTEMS (PROJECT MANAGEMENT)

ADVISE PROGRAMMERS AND SYSTEMS ANALYSTS ON DIVISION'S POLICY AND PROCEDURES FOR DATA PROCESSING.

ASSIGN ANALYSTS TO PREPARE A PROPOSAL FOR PERSENTATION TO A REVIEW BOARD.

ASSIGN PERSONNEL TO PROJECT QUALITY REVIEW BOARDS (QRBS).

ASSIGN PERSONNEL TO PROJECTS.

ASSIGN SYSTEMS AND PROGRAMMING PROJECTS TO DATA PROCESSING SUPPORT TEAMS.

ASSIGN TASKS TO PROGRAMMER.

DEFINE THE FUNCTIONS TO BE ASSIGNED TO PROGRAMMERS.

DETERMINE THE LEVEL OF PERSONNEL NECESSARY TO DEVELOP THE SYSTEM.

DETERMINE TIME ESTIMATES TO COMPLETE TASKS.

DEVELOP WORK SCHEDULES FOR PROJECT.

PARTICIPATE IN WALKTHROUGHS AND REVIEWS OF PROJECTS AND/OR SYSTEMS.

REVIEW A PROPOSED PROJECT WITH SYSTEMS ANALYST AND USER TO DEFINE THE SCOPE OF THE PROJECT.

REVIEW SYSTEMS DEVELOPMENT PLANS TO DETERMINE WHAT ADDITIONAL STAFF AND RESOURCES ARE NECESSARY.

DESIGNING SYSTEMS (SYSTEM PROCEDURES)

ANALYZE AND DETERMINE THE TYPE OF FORMS REQUIRED TO COLLECT DATA.

ANALYZE THE LOGIC FUNCTIONS REQUIRED IN THE SPECIFICATIONS FOR WRITING PROGRAMS.

CONDUCT HARDWARE CAPACITY PLANNING REVIEWS.

DETERMINE THE TYPES OF EQUIPMENT REQUIRED TO SUPPORT A SYSTEM.

DETERMINE THE [illegible] OF PROGRAMS AND ROUTINES THAT WILL BE REQUIRED TO DEVELOP A SYSTEM.

PREPARE WORKPLANS FOR PROJECT.

REVIEW AND EVALUATE DATA AND INFORMATION COLLECTION PROCEDURES.

REVIEW AND EVALUATE SPECIFICATIONS FOR A PROPOSED SYSTEM.

REVIEW FINDINGS OF FEASIBILITY STUDIES TO DETERMINE APPROPRIATE ACTIONS.

REVIEW MANUALS FOR [illegible] SOFTWARE.

[illegible] AND [illegible] PRIORITIES FOR PROJECT START-UPS.

[illegible] ADEQUACY OF APPROACH.

REFERENCES

1. Mayo, Clyde C., "Three Studies of Job Inventory Procedures: Selecting Duty Categories, Interviewing, and Sampling," Lackland AFB, Tx.: Personnel Research Division, Air Force Human Resources Laboratory, November, 1969.

2. McCormick, E.J. and Ammerman, H.L., "Development of worker Activity Check Lists for Use in Occupational Analysis," WADD-TR-60-77, Lackland AFB, Tx.: Personnel Laboratory, July, 1960.

Corman, Steven, Annandale, Virginia. (Wed. A.M.)

Computerized Optimal Weight Scoring (COWS): A Comparison of Three Procedures

This paper documents the results of a simulation study comparing three computerized optimal weight scoring (COWS) procedures based on item response theory, with conventional scoring. The effect of the COWS procedures is to differentially weight test items as a function of examinee ability and the item characteristics. Low ability examinees are given very low weights on difficult test items, lowering the effects of guessing, and decreasing test error.

A MONTE CARLO COMPARISON OF THREE OPTIMAL TEST SCORING PROCEDURES

Steven Gorman
Electronic Data Systems , aley . ta

Differentially weighting items to improve the psychometric properties of multiple-choice aptitude or achievement test scores has long been of theoretical and practical concern. Applications of item weighting procedures which were empirically developed or based on classical test theory did not generate the desired psychometric properties (Bejar & Weiss, 1973; Downey, 1976).

Lord (1968) successfully applied item weighting based on the three parameter logistic model developed by Birnbaum (1968) to the Verbal Scholastic Aptitude Test (VSAT). The effect of this procedure is to differentially weight test items as a function of examinee ability on the basis of their item discriminatory power, difficulty, and susceptibility to guessing. Accordingly, low ability examinees will be given very low weights on difficult test items, thus lowering the effects of guessing, and decreasing test error. The use of optimal item weights eliminates more test score error at low ability ranges, where guessing on difficult items is more prevalent. Because of the difficulty (until recently) of accurately estimating these item parameters, and the computations involved in using Lord's procedure, this optimal item weighting method has not been widely adopted. The accurate estimation of the three parameters of test items has become commonplace with the development of several new computer programs, namely LOGIST (Wood, Wingersky, & Lord, 1976), ANCILLES (Croll & Urry, in preparation) and FIT3IP (Gugel, in preparation). Three FORTRAN computer programs, MAXLIKE, BAZEMODAL, and OWENSTAT have been developed to estimate ability based on the use of optimal scoring weights.

The present study simulates a live testing situation by introducing item parameter estimation error. It compares the performance of conventional unit weight scoring with three optimal item weighting methods using the item parameters estimated by ANCILLES. The data in this investigation were produced with a Monte Carlo procedure which generated examinee item responses.

TECHNIQUES INVESTIGATED

The psychometric characteristics of test scores for three optimal item weight scoring procedures, and a unit weighting scoring method were investigated using three idealized types of test distributions.

The three optimal scoring procedures are based on Birnbaum's three parameter logistic model which states that the probability of a correct response given an ability level is:

$$P(u_i = 1|\theta) = c_i + (1 - c_i)[1 + \exp(-1.7a_i(\theta - b_i))]^{-1} \qquad (1)$$

where a_i is the item discriminatory power, b_i is the item difficulty, and c_i is the item coefficient of guessing.

The three optimal scoring procedures, and a conventional scoring procedure used in the study are described in the following paragraphs.

Owen Bayesian Scoring

This method is described in Gorman (1979) and based on Owen (1975). The method uses the Owen algorithm scoring procedure in a sequential rather than adaptive mode, in the computer program OWENSTAT.

Bayes Modal Scoring

This procedure is based on the work on Samejima (1969) and assumes that ability is normally distributed. The ability estimate is the maximum value for:

$$B_v(\Theta) = N(\Theta) \prod_{i=1}^{n} P^{u}i(\Theta)Q^{u}i(\Theta) \qquad (2)$$

where $N(\Theta)$ is the normal Gaussian distribution, $P^{u}i(\Theta)$ is the probability of a correct response to item i given ability Θ, and $Q^{u}i(\Theta) = 1-P^{u}i(\Theta)$.

Maximum Likelihood Scoring

This procedure, described in Lord & Novick (1968), states that the ability estimate is that value which maximizes $L_v(\Theta)$ in:

$$L_v(\Theta) = \prod_{i=1}^{n} P^{u}i(\Theta)Q^{u}i(\Theta) \qquad (3)$$

where $L_v(\Theta)$ is the maximum likelihood ability estimate.

The above two scoring equations were solved in the computer programs BAZEMODAL and MAXLIKE by use of a modified Newton-Raphson algorithm.

Z-Score Transformation

This is a conventional unit weighting scoring procedure where:

$$Z = \frac{X-\overline{X}}{S.D.} \qquad (4)$$

where X is the examinee's raw test score, $\bar{X}$ is the average raw test score, S.D. is the standard deviation of raw scores, and Z is the ability estimate.

METHODOLOGY

Ideal tests were constructed consisting of two levels of item quality, low (a_i = 0.8) and high (a_i = 1.6). Three test types were developed which varied the item difficulty (b_i) distributions in order to provide maximum test information (Birnbaum, 1968) at:

(1) evenly distributed values over the ability range Θ = -2.5 to +2.5 (rectangular)

(2) midpoints of even-sized areas of the Gaussian distribution (normal) and

(3) at the ability mean (peaked).

The susceptibility to guessing, c_i was set for all test items at .15, a reasonable average for a five alternative multiple choice test (based on Jensema, 1976).

Data were generated in accordance with Birnbaum's (1968) three parameter logistic model (equation 1) in two parts:

(1) creating examinee responses for the estimation of item parameters, and
(2) creating examinee responses to all test items with known item parameters.

When the item parameters are specified, the probability of a correct response is strictly a function of the simulated examinee's (sims) ability. To generate dichotomous responses, the probability of a correct response is determined for the examinee's ability from equation 1. If this probability is greater than or equal to a random number between 0 and 1, then a correct response is generated, else an incorrect response is generated. The item parameters for these items along with items from a separate study (Gorman, 1980) were then estimated by ANCILLES based upon the administration of tests of 51 item length to 2000 sims representing a normal population. These estimated item parameters were used in the three optimal scoring procedures.

STUDY 1

In Study 1, a group of 500 normally distributed sims, generated by the LLRANDOM computer program (Learmonth & Lewis, 1973), were administered all tests. The criterion evaluated in Study 1 was the fidelity coefficient, the correlation of known ability with ability estimated by each scoring method on the 20 and 30 item tests. This statistic has been widely used in other testing research. Samejima (1977) decries the use of this statistic, stating that since it is a correlation coefficient, it is dependent not only upon the test, but also upon the specific group of examinees tested. Testing the same sims from a normally distributed population will hold constant across test scoring methods the effect of the ability distribution of examinees. It should be demonstrable that a theoretical test can have a high fidelity coefficient, yet have poorer psychometric properties as a function of ability than other theoretical tests.

RESULTS

The results of Study 1 are displayed in Table 1. The fidelity coefficients, the correlations between known and estimated ability, are typically greater for tests of all types, lengths, and item qualities scored with the optimal scoring methods. The only exception is with the peaked high item quality tests, where the conventional scoring procedure ordered examinees better, on average, than with the maximum likelihood procedure. On the low item quality tests, the ordering of examinees is greatest with the peaked tests. On the high item quality test, the normal tests ordered examinees best. Note that the 20 item normal test with higher item quality scored with any of the optimal scoring procedures has higher fidelity coefficients than the 30 item conventionally scored test. Also note that as test length and item quality increase, the fidelity coefficients of the optimal scoring procedures become much greater than those of the conventional method.

TABLE 1

Correlations between Known Ability and Ability Estimated by Four Scoring Methods on Three Types of Test Distributions*

20 ITEM TEST

Test Type	Low a_i R	Low a_i P	Low a_i N	High a_i R	High a_i P	High a_i N
Method						
BAZEMODAL	842	891	877	927	906	943
MAXLIKE	839	888	874	926	876	941
OWENSTAT	840	892	875	922	911	939
Z-SCORE	828	888	869	908	904	929
30 ITEM TEST						
BAZEMODAL	907	923	923	949	926	954
MAXLIKE	905	921	925	949	891	954
OWENSTAT	905	924	922	944	924	949
Z-SCORE	894	919	918	933	912	938

*Decimals omitted

DISCUSSION

With low item quality, the 20 item peaked test and the 30 item peaked and normal tests provided the greatest fidelity coefficients, the correlation between the ability estimates and known ability. With high item quality, the normal tests have the greatest fidelity coefficient. This follows from item response theory, which holds that item information becomes more leptokurtic as item discriminatory power (a_i) increases. Thus, since item information is additive, a peaked test with low a_i values should differentiate examinees over a much broader range than a peaked test with high a_i values.

This study has shown that on this one criterion, the non-rectangular tests are better average measures of ability. Study 2 will examine two other psychometric properties of these test scoring methods.

STUDY 2

In Study 2, the sample consisted of 100 sims at each of 11 evenly spaced ability values on the ability continuum -2.5 to +2.5. The instruments used were the three types of 30 item tests. These sims provided data to compute statistics as a function of examinee ability. The criteria evaluated are score bias and test score precision. Test score bias is the average difference between the known examinee ability and the ability estimated by each scoring method. Test score precision is given by the test score information value, an indicator of the usefulness of the test scores for differentiating ability at that ability level. Test score information is inversely related to the square of the standard error of the ability estimate, and varies as a function of ability.

RESULTS

The mean difference between known ability and the ability estimated by the four procedures was computed for the 100 sims at each of the 11 ability levels. For ease of comparison, these data were aggregated by computing an average of the absolute values of the 11 score bias figures. These values are located in Table 2. The data show that the bias is lower (excepting the Bayes modal scored peaked test of low item quality) for all optimal scoring methods for all test types and levels of item quality than with conventional scoring. The bias is most severe with the peaked test, and this bias increases with higher item discriminatory power.

TABLE 2

Average Absolute Value of Score Bias for Ability Estimated by Four Scoring Methods on Three Types of Test Distributions*

	Low a_i			High a_i		
Test Type	R	P	N	R	P	N
Method						
BAZEMODAI	18	27	18	13	38	14
MAXLIKE	11	17	09	12	25	12
OWENSTAT	11	17	10	15	30	14
Z-SCORE	20	26	21	13	44	22

*Decimals omitted

Two criteria need to be measured when evaluating test score precision. One is the average precision (computed by estimated test score information), and the other is the equiprecision over the ability range. Since the items for all types of test distributions have the same item discriminatory power and susceptibility to guessing, all tests share identical potential test score information. The test score information could only be measured over the range -2.0 to +2.0 with this research design, thus some of the test score information for the non-peaked tests is not measured. Therefore, average precision, listed in Table 3, should only be reviewed within test types. For the tests consisting of items with low item discriminatory power, all scoring procedures yielded roughly the same average information. However, with the tests of higher quality items, the optimal scoring procedures provided greater information than with conventional scoring, except with the peaked test.

TABLE 3

Average Test Score Information for Four Scoring Methods on Three Types of Test Distributions

	Low a_i			High a_i		
Test Type	R	P	N	R	P	N
Method						
BAZEMODAI	4.47	5.64	4.99	9.85	10.72	12.47
MAXLIKE	4.39	5.60	4.83	9.62	11.04	12.61
OWENSTAT	4.49	5.57	4.97	9.32	10.45	12.08
Z-SCORE	4.16	5.79	4.86	7.44	12.08	10.25

The equiprecision of the tests by scoring method is measured by the coefficient of variation (CV) of test score information, listed in Table 4. The greater this CV value, the less equiprecise the test score information. The peaked tests provide the greatest CV values, with the optimal scoring procedures yielding greater values than with conventional scoring. The peakedness of information increases, as expected, with the higher item discriminatory power items. The rectangular tests give more even test score precision, with this phenomenon more evident on the higher item discriminatory power test. Although the conventional scoring of the rectangular high item quality test has lower CV values than with optimal scoring, the three optimal scoring methods all yielded greater information at all nine ability levels than with conventional scoring.

TABLE 4

Coefficient of Variation of Test Score Information for Four Scoring Methods on Three Types of Test Distributions

	Low a_i			High a_i		
Test Type	R	P	N	R	P	N
Method						
BAZEMODAL	24	59	25	29	123	53
MAXLIKE	25	63	28	18	136	53
OWENSTAT	27	60	24	22	128	49
Z-SCORE	29	52	36	16	94	52

DISCUSSION

These optimal scoring procedures weight items as a function of their item information (Birnbaum, 1968), the contribution of each item to decrease test score error at each ability level. The item information is a function of the item's a_i, b_i, and c_i parameters. Since the a_i and c_i values are fixed in each test, the full capacity of these scoring procedures is not being demonstrated. Both studies only show the capacity of these procedures to effectively weight items as a function of their appropriateness in difficulty relative to the examinee's ability, not their capacity to weight items as item discriminatory power and item coefficient of guessing vary. In spite of this shortcoming, the optimal scoring procedures show a significant increase in their ability to successfully order the ability of examinees relative to the conventional scoring procedure. The optimal scoring procedures also measure examinees with more precision and less bias than the conventional means.

This study also assumes that the multiple choice test has five response alternatives. For tests with only four item choices, or where the chance of successful guessing is greater that .15, the scoring properties would tend to diminish. The optimal scoring procedure properties would likely diminish slightly, while the conventional scoring properties would drop more significantly. This is due to the optimal weighting procedures capacity to diminish the effect of test score error due to guessing, while conventional scoring procedures are less capable of reducing the effects of guessing.

CONCLUSION

In this Monte Carlo study, three optimal test scoring procedures using estimated item parameters provided better psychometric properties of ability estimates than the conventional procedure. The optimal scoring procedures provided the greatest advantage over conventional scoring on rectangular tests composed of items with high discriminatory power. Test publishers should seriously consider using one of these optimal methods to score their multiple choice examinations. With the ready availability of computer programs to estimate item parameters and optimally score tests, the benefits of enhanced measurement of ability should outweigh the slight increase in computer costs. This study also showed that the fidelity coefficient criterion is only a group psychometric indicator, and does not show the capacity of the test to measure low and high ability examinees. This criterion should be used with caution to avoid making erroneous conclusions.

REFERENCES

Bejar, I.I. and Weiss, D.J. Comparison of four empirical differential item scoring procedures. Research Report 73-2, Psychometric Methods Program, Department of Psychology, University of Minnesota, 1973.

Birnbaum, A. Part 5. Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. and Novick, M.R. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.

Croll, P. and Urry, V.W. ANCILLES: A program for estimation of the item parameters of normal ogive and logistic mental test models.

Downey, R.G. Item option weighting of achievement tests: a comparative study of methods. Paper presented at the American Psychological Association Convention, Washington, D.C., Sept., 1976.

Gorman, S. A comparison of Bayesian adaptive and static tests using a correction for regression. Paper presented at the third Biennial Conference on Computerized Adaptive Testing, Wayzata, Minnesota, 1979.

Gorman, S. A comparative evaluation of two Bayesian adaptive ability estimation procedures with a conventional test strategy. Doctoral dissertation. Catholic University, 1980.

Gugel, J. Estimating the parameters of the normal ogive three parameter model via an extension of theory and methodology of Kelly and of Lord. (in preparation).

Jensema, C.J. A simple technique for estimating latent trait mental test parameters. Educational and Psychological Measurement. 36, 705-715, 1976.

Learmonth, G.E. and Lewis, P.A.W. Naval Postgraduate School Random Number Generator Package: LLRANDOM. Research Report NPS55LW73061A, Naval Postgraduate School, Monterey, California, 1973.

Lord, F.M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 28, 989-1020, 1968.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph No. 17, 1969.

Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977, 1, pp. 233-247.

Wood, R.L., Wingersky, M.S., and Lord, F.M. LOGIST: A computer program for estimating ability and item characteristic curve parameters. Research Memorandum 76-6, Princeton, N.J. Educational Testing Service, 1976.

AD P001324

AIR FORCE CIVILIAN PROMOTION APPRAISAL SYSTEM DEVELOPMENT

R. Bruce Gould, PhD
Chief, Performance Evaluation Section
Air Force Human Resources Laboratory

A system for ranking civilian promotion eligibles was developed to meet legal and operational requirements. Demographic data and supervisor and peer ratings were obtained on a 20,000 case target sample to establish a pool of potential ranking elements. Variable reduction procedures eliminated those elements which were redundant, showed greatest potential for adverse impact, or were less defensible legally or psychometrically. The remaining variables were submitted to 12-member promotion data panels for each of 23 homogeneous job family clusters. Policy capturing techniques resulted in regression weighted promotion algorithms which use between 5 and 18 behavioral dimensions, depending on the job family and supervisory or nonsupervisory classification of the job, to rank promotion candidates. Correlations between promotion panel rankings and algorithm predicted rankings ranged from .89 to .96 indicating consistent panel member policies and high interrater agreement. Separate reliability panels convened for two job families resulted in correlations of .94 to .98 between their rankings and those of the algorithm development panels. A single supervisory rating form and 40 job-type specific algorithms have been recommended for operational implementation.

The Air Force Human Resources Laboratory began development of the Civilian Promotion Appraisal System (CPAS) in July 1977 at the request of the Directorate of Civilian Personnel. The system was to provide a rating process for determining promotion potential and rank ordering all General Schedule (GS) and Federal Wage Scale promotion eligibles. The rating system would be the last ranking stage in the Air Force Promotions Placement and Referral System (PPRS). The earlier stages screen candidates on basic eligibility, e.g., satisfying minimum skills, knowledge, and experience requirements established by a job analysis of each specific vacant position. CPAS was to restore credibility to the civilian promotion appraisal process, be operationally efficient, and manage rating inflation.

The research approach was designed to meet constraints set by the user and was subsequently found to meet later constraints imposed on promotion systems by legal and higher authority directives. In 1978, the Office of Personnel Management (OPM) directed that promotion systems only use those aspects of current job performance ratings which are related to the specific new job in promotion actions. Public Law 95-454, otherwise known as the Civil Service Reform Act of

1978, stated that task-level performance measures must be a basis for promotion actions. The Equal Employment Opportunity (EEO) and Uniform Guidelines for Employee Selection (1978) dealt with adverse impact issues and specified selection system developmental actions which must be taken to minimize adverse impact while maximizing validity. These constraints plus a review of case law (Cascio & Bernardin, 1981) mandate that the new promotion system must not rely on subjective measures of workers' affective states, e.g., traits or characteristics, or must not use overall job performance measures since they contain performance elements not demonstrated to be associated with the new job. Likewise, the performance potential measures could not ask raters to project performances of job aspects which were beyond those observable in the current jobs.

Legal defensibility of the promotion system was a major goal of the project. Under the Uniform Guidelines (1978), any system which results in adverse impact must have fulfilled several developmental requirements to absolve the user from liability. The probability that any major selection system will have adverse impact is sufficiently high that all such systems should be developed adhering to the guidelines. To satisfy the guidelines, all possible selection components must have been considered and systematically selected to provide a system which maximizes objectivity and validity while keeping to a minimum elements which will contribute to adverse impact. Elements showing the <u>least</u> amount of bias must be selected but not to the point of <u>sacrificing</u> validity. Where criterion validity is not available, construct and content validity should be used to develop the system and criterion validation should become an ongoing process during system use.

This report outlines the procedures used to develop the experimental data collection devices, organize the civilian jobs into a manageable number of job families, collect the data, select the variables for inclusion in the final promotion system, develop the operational algorithms or formulas for ranking candidates, and recommends steps for managing rating inflation.

Data Collection Instruments

Three experimental data collection instruments were developed. The instruments contained items for research purposes as well as potential promotion ranking elements. Literature surveys, comparisons of existing Government and industry rating forms, job analyses conducted in similar military occupations, and experimenter subjective judgments provided the pool of potential promotion ranking elements from which a demographic questionnaire and a rating form were developed. The 102 item Demographic Questionnaire captured traditional demographic information such as age, sex, and ethnic background; promotion, education, and job training history; time in service, grade, and on-the-job measures; and interest in and attitudes toward promotion incentives such as educational activities the employee was willing to engage in to become promotion eligible. The 100 item Worker Characteristic Rating Booklet used 9-point adjectival anchored rating scales to obtain ratings on a wide range of operationally defined worker job performance measures (skills and knowledges). Also included were ratings of overall job performance and predicted performance in the next higher job and scores on paper-and-pencil measures of verbal, quantitative, mechanical, electrical, and spacial aptitude. These later measures were included to permit an evaluation of rater accuracy.

The third instrument was the Civilian Personnel Examination (CPE) which was designed to measure the workers' quantitative, verbal, spacial, mechanical, electrical, and administrative aptitudes. The CPE was a revised version of the Air Force Qualifying Examination, Form J (Vitola, Massey, & Wilbourn, 1971). The CPE was not a candidate component for the promotion system but again provided a means for assessing rater accuracy and determining the reliability of specific skills and knowledge ratings.

Job Family Specifications

To obtain a manageable number of logical and homogeneous clusters of job types, the some 1,500 Air Force civilian job series were combined into 23 job families. Eight position classifiers and research psychologists were divided into two independent panels and instructed to arrange the 1,500 job series into some 20 homogeneous clusters based on task subject matter and job requirements. Their results were then compared and differences resolved. There was near perfect overlap between their clusters. The joint session resulted in 23 job family groupings. There were three professional, eight technical, four administrative, two clerical, and six trades and crafts job families.

Data Collection

A stratified random sample of 20,000 target workers from the 784 most populous CONUS bases were selected. The sample was stratified by grade, series, sex, and ethnic category with 800 to 1,000 cases representing each of the 23 job families. The sample contained a greatly overrepresentative number of minority members so that cell sizes would be adequate for bias and adverse impact analyses. The Demographic Questionnaire, with self-addressed envelopes to maintain privacy, were sent to 20,000 target workers. Each of their supervisors received a Worker Characteristics Rating Booklet as did one of each target worker's peers.

Statements outlining the "Research Purposes Only" use of the ratings and the nature and purpose of the entire research project, and the efforts of local project monitors are credited with the enthusiastic support of the participants. Seventy-one percent of the materials were returned with 65% of the target workers (N = 12,865) having complete matched and usable data sets. The substantial variance in the behavioral performance element ratings, constructive write-in comments, and completeness of the data attest to the support workers and supervisors gave to the promotion system development effort.

The CPE was administered to 2,000 of the 20,000 target workers to obtain aptitude measures for the rater accuracy studies. Worker characteristics ratings were obtained from peers to aid in these analyses and provide criterion measures. The rater accuracy results go beyond the purpose of this paper except to say that supervisors and peers were very effective in predicting performances on aptitude tests and there were no significant or practical differences between their rating effectiveness.

Variable Selection

To reduce the 210 demographic and rating variables to a workable subset, a large data matrix was constructed for rating each of the elements on a series of selection criteria. Four basic types of criteria were used: (1) validity, (2) uniqueness, (3) bias, and (4) legal defensibility. Correlations of the variables with peer ratings of overall performance and projected performance at the next higher levels served as estimates of item validity. Comparisons of mean values by race and sex category were used to evaluate item bias. Intercorrelations between items were used to identify items capturing unique and redundant variance. Thus objective measures of each item's validity, uniqueness, and bias were available. Subjective ratings of legal defensibility were made by personnel specialists and research psychologists knowledgeable in case law.

A panel of eight research psychologists and two personnel specialists was convened to select the workable subset of variables from the data matrix. Their goal was to select items which were valid, minimized bias and hence potential adverse impact, contributed unique variance, and had not been demonstrated to be legally indefensible. First, all items which were not legally defensible were deleted. This limited the pool to observable behaviors and deleted demographic variables which are highly related to past opportunities and associated with past discriminations. Then judgments were made to select the final subset of 24 rating elements. There were three overall ratings (job performance, supervisory, managerial); 12 behavioral ratings such as responsiveness to directions, and self-sufficiency; three ability ratings (quantitative, reading, data interpretation); four motivation indicators (productivity, initiative, speed of completion, and amount of working time spent in productive efforts); and two composite variables which combined behavioral or motivational measures. The above categories are used for summary purposes only. In the actual rating elements, only operational definitions are used which describe observable performances which are related to say managerial behavior, rather than using the words "managerial performance." By this procedure, individuals who are not managers can still be rated on observable behaviors which are related to managerial performance.

Promotion Algorithms

Panels of 11 to 12 subject matter specialists (SMSs) were convened for each of the 23 job families to select the final set of factors to be used for rating and ranking promotion eligibles in their job family.

For each panel, stratified random sampling techniques were used to balance representation by race (Black, Hispanic, other), sex, and supervisory/nonsupervisory status. Selectees proportionally represented job series, locality, and major commands while meeting minimum experience and availability requirements.

Each panel's first task was to review and discuss the 24 prospective rating elements. Then to select a set of 6 to 10 elements which were observable on-the-job in their job family and were related to ability to perform at the next level nonsupervisory job. A second set was selected for supervisory jobs. Data decks were then produced which contained only the selected rating elements. The data pertained to workers in the original 20,000 cases who were members of the specific job family.

The panel's second task was to sit as a promotion board using procedures similar to those used by military boards. Members were given decks of 70 to 80 cases and asked to rank-order the cases according to their promotability or potential to perform in the next higher-level job. Decks were labeled according to specific grade level and supervisory/nonsupervisory nature of the candidate job. Each member received 9 to 15 data decks to rank order. Wage grade job families considered only nonsupervisory jobs while the one wage supervisory family considered only supervisory jobs. This accounts for the small number of decks considered by some job families. Unknown to the participants, one duplicate deck was included so inconsistent raters could be identified. An additional two panels were convened to replicate, and cross-validate, the results of two of the job family panels.

Judgment analysis techniques (Christal, 1963) were used to capture the promotion policy of each individual and the aggregate policy of each panel. Of 277 SMSs, only two were removed from the analyses because of inconsistent within-rater policies. Judgment analysis techniques use the available data elements (6 to 10) as predictors and each case's rank order as the criterion to obtain regression weights and multiple Rs. The sum of the regression weights times the respective ratings all added to a regression constant yields each case's predicted rank. The correlation (R) between the predicted and actual ranks indicates the efficiency with which a panel's policy was captured.

Multiple Rs ranged from .89 to .96 indicating the substantial predictive efficiency with which the regression weights captured each panel's promotion policy. These values are particularly noteworthy considering that any inconsistency within raters and differences between panel raters introduces error. There were no within job family grade-level differences in rating policies, but there were significant differences between all supervisory and nonsupervisory rating policies. Forty algorithms were developed, 18 supervisory and 22 nonsupervisory. The two cross-validation panels resulted in correlations of .94 to .98 between their rankings and those of the algorithm development panels.

Regression weights are difficult to interpret or explain to users promotion candidates, so alternative weighting systems w investigated. A unit weighting scheme coding raw score weights l than .5 as "0" and all other weights as "1" produced virtua identical results and was adopted for operational use. Some of rating elements received "0" weights in all algorithms and the numl of operational behavioral dimensions was reduced to 19.

Operational Rating Form

A single universal operational rating form was recommended for use f all GS and Federal Wage Scale employees. All 19 remaining behavior dimensions contain operationally defined components which can be rat for all employees. Operationally, only those dimensions which a unit weighted will be used in a specific promotion ranking. T universal form satisfies the user's request for simplicity eliminating the current multiple forms and removing the necessity f multiple rating forms for individuals being considered for positio outside their current series. As will be explained, the single fo is a necessary requirement to permit inflation management.

Inflation Management

Four specific recommendations were made to permit inflation manageme of the ratings. These recommendations concern the rating scal rating all individuals in a unit at the same time using a univers form, providing a separation between job performance and promoti rating systems, and providing a feedback system. The rating scale i a 9-point behaviorally anchored scale where the rated individual i compared to the average employee. The average employee is defined a being highly effective and motivated in the performance of his/he job. Ratings using the top or bottom two scale points must b substantiated by citing specific performance events.

By rating all individuals in a unit simultaneously, raters can mak better comparisons among subordinates. Each supervisor's ratings ar to be summarized on work sheets and the summaries submitted with th ratings to an indorsing official. Indorsing officials can then chec for leniency, halo, or response set type rating errors. Further, th summaries permit the indorsers to make comparisons across raters an ensure that ratings are consistent with each organization unit's productivity. Indorsers will be authorized to andate that ratings be consistent with actual performance.

The greater the number of uses for a rating, the greater the pressure on the supervisor to inflate the rating. For this reason the CPAS ratings are distinct in time and purpose from the job performance ratings rendered under the General Manager Appraisal System (GMAS) and Job Performance Appraisal System (JPAS). The GMAS and JPAS purposes and procedures were presented earlier at this conference (see Thompson, Cowan, & Guerrieri, 1981). The only relationships between the systems are that individuals must have at least satisfactory GMAS and JPAS ratings to be promotion eligible and ability to render appropriate CPAS ratings is a component of each supervisor's GMAS or JPAS rating.

Rating feedback systems usually provide mean summary values for comparisons with peers and comparing subordinate units rating policies. Anecdotal information suggests that organizations with lower mean ratings feel they may have been unfair to their employees and increase the average of their next ratings. The feedback information actually increases rather than decreases inflation thus creating an inflation pump. With CPAS, a feedback system is still recommended. However, the recommended system provides only variances in ratings, not mean values. Accompanying guidance will point out that "the more variance the better." Lack of variance is the result of poor rating management and inability to differentiate performance within individuals and between individuals. Inflation reduces score dispersion hence variance. This variance yardstick is also to be used to guide subsequent performance ratings of supervisors. The job performance rating systems mandated by CSRA-78, and as developed by the Air Force, required that all supervisors have a critical supervisory ability job element. Therefore, all supervisors are rated on their ability to rate. Variance feedback will make supervisors accountable for their rating styles.

Summary

CPAS developers have attempted to provide the Air Force with a civilian promotion rating system which is creditable, valid, legally defensible, and manageable. The system will be implemented by April 1982 after users and participants have had significant training on the system goals and mechanics to include training raters how to rate.

REFERENCES

Casio, W.F., & Bernardin, H.J. Court cases relevant to employment decisions: Annotated bibliography. AFHRL-TR-80-44, AD-B055 755L. Brooks Air Force Base, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory, February 1981.

Christal, R.E. JAN: A technique for analyzing group judgment. PRL-TDR-63-3, AD-403 813. Lackland Air Force Base, TX: 6570th Personnel Research Laboratory, Aerospace Medical Division, February 1963.

Thompson, N.A., Cowan, D.K., & Guerrieri, J.A. The Air Force job performance appraisal system. Paper presented to the 23rd annual conference of the Military Testing Association. Arlington, VA: U.S. Army, October 1981.

Uniform Guidelines for Employee Selection Procedures. From Federal Register 43(166), August 15, 1978.

Vitola, B.M., Massey, I.H., & Wilbourn, J.M. Development and standardization of the Airman Qualifying examination - Form J. AFHRL-TR-71-28, AD-730 592. Lackland Air Force Base, TX: Personnel Division, Air Force Human Resources Laboratory, May 1971.

AD P001325

EXPLORATION OF INSTRUCTIONAL APPLICATION OF VIDEODISC TECHNOLOGY IN THE MARINE CORPS

H. William Greenup
Marine Corps Development and Education Command

Steven Levin
Interactive Television Company

SUMMARY

The Education Center, Marine Corps Development and Education Command, is experimenting with the instructional applications of a videodisc-based interactive television system. Videodisc technology offers the potential for a wide range of military training applications including archival storage and information retrieval, spatial data management, tactical mapping, multi-aspect filming, and nonresident instruction. The Marine Corps, under a contract with Interactive Television Company, Inc., has produced an initial disc that presents a tactical map of its combined arms training area at Twentynine Palms, California. This paper discusses the instructional and technical aspects of videodisc technology involved in the development of such a program. The authors describe the hardware and programming techniques used in the production and presentation of interactive instructional programs which are appropriate to the learner's responses and needs. They also review some of the ways interactive television already is being used in education and training and project future developments in interactive video instruction.

INTRODUCTION

We may be reaching the productivity limits of our current training technology. Massive efforts to improve "stand-up" lectures, printed training materials, and "hands-on" laboratory experience have begun to yield too few significant returns. We need revolutionary new techniques that will break through the constraints imposed by existing training technology and allow us to beat the current tradeoffs that must be made among costs, quantity and quality. Videodisc technology is one of the most promising sources of these new techniques.

Videodisc technology is creating opportunities for new kinds and forms of training devices at surprisingly low cost.

At the heart of this technology is the capability to access tens of thousands of color images, including stereo sound, in seconds or fractions of seconds. When coupled with the ubiquitous microprocessor, videodisc technology offers new opportunities for training applications as well as new issues for the training specialist.

OPTICAL VIDEODISC TECHNOLOGY

A videodisc is similar to an audio record except that each side of a 12-inch disc contains 30 minutes of television. Since television is the rapid presentation of pictures at a rate of 30 images per second, each side of a videodisc contains 54,000 still pictures in color.

The player for a videodisc is much like a turntable except that under computer control the player can rapidly find any particular part of the videodisc. Specifically, any one of the 54,000 images, or any part of the 30 minutes of television on a videodisc, can be located typically in a fraction of a second. A videodisc provides a combination of moving and still images, any part of which can be quickly located.

Videodisc are made like audio records using original materials that can be movies, videotapes, pages of text, tables of numbers, graphs, charts, maps, drawings, diagrams, or photographs. From the original materials, a master disc is produced that in turn is used to "press" multiple copies speedily and at low cost.

Three types of videodiscs and videodisc players are now available. The players are characterized by the market they are targeted for: general consumer usage and industrial/educational applications. Various potential manufacturers, such as JVC, IBM, Xerox, Zenith, to name a few, are waiting in the wings, but only six are now marketing videodisc players.

- Magnavox, a wholly owned subsidiary of Phillips, began selling a consumer model player for about $800 in December 1978.

- Discovision (DVA), under license by MCA, began general sales of an industrial player for about $3,000 in June 1979.

- Thompson-CSF began sales of an industrial player for about $3,500 early in 1980.

- Pioneer began sales of a consumer player for about $700 late in 1980.

- RCA began sales of their consumer videodisc player early in 1981.

- Sony began limited sales of an industrial videodisc player in mid-1981.

The difference between consumer and industrial systems are significant. For instructional settings, the industrial players incorporate microprocessor control for fast random access, pre-programmed branching, and frame selection. Microprocessors also have been installed in the less expensive consumer players to provide random access, branching, and frame selection. However, the servo-mechanism in these players requires from 15 to 20 seconds to locate a frame in the worst case. Maximum access time for the industrial players is under four seconds. The consumer players lend themselves to instructional application, but the functional capabilities of the industrial players may be sufficiently greater as to justify their higher cost.

With the exception of the RCA system, the videodisc players discussed above are all optical, laser-based systems. They use a 12-inch disc with a spiral track. The track is pitted with oblong depressions or micropits about 1 micron deep that vary in accordance with the audio or video information they represent. During playback the disc spins at 1,800 revolutions per minute while these micropits modulate a low power helium-neon laser focused on the track and thereby generate a signal that is processed and passed to a standard video monitor (i.e., to the antenna terminals of a TV set).

One video frame is stored on each track and there are 54,000 tracks per disc. Video, audio, and still photographic information can all be intermingled on these discs. These videodisc systems effectively provide rapid access to 30 minutes of video information, 30 minutes of analogue audio information, 54,000 still photographs, well in excess of 400 hours of digitized audio information, 30 minutes of motion picture information, or various combinations of these media. The point to be made is that videodiscs provide rapid random access to a lot of information which can be inexpensively stored and replicated.

APPLICATIONS OF VIDEODISC TECHNOLOGY TO TRAINING

One of the most exciting aspect of videodisc technology is that it makes possible an entirely new set of training experiences, ranging from new kinds of training movies to low-cost simulators.

Interactive Movies

Interactive Movies translate movie viewing into an active participatory process. In effect, the viewer becomes the director and controls many features of the movie. A sampling of feature controls available to the viewer is the following:

1. Perspective. The movie can be seen from different directions. In effect, the viewer can "walk around" ongoing action in the movie or view it from above or below.

2. Detail. The viewer can "zoom in" to see selected, detailed aspects of the ongoing action or can "back off" to gain more perspective on the action and simultaneous activity elsewhere.

3. Level of Instruction. In some cases, the ongoing action may be too rich in detail or it may include too much irrelevant detail. The viewer can hear more or less about the ongoing process by so instructing the Interactive Movie System.

4. Level of Abstraction. In some instances the viewer may wish to see the process being described in an entirely different form. For example, the viewer might choose to see an animated line drawing of an engine's operation to get a clearer understanding of what is going on. In some cases, elements being shown in the line drawings may be invisible in the ongoing action -- for instance, electrons or force fields can be shown.

5. Speed. Viewers can, of course, view the ongoing action at a wide continuous range of speed -- including reverse action and no action (still frame).

6. Plot. Viewers can change the "plot" to see the results of different decisions made at selected times during the movie.

A typical application for Interactive Movies would be in training (and aiding) equipment technicians. The technician could not only see how a particular part is located and installed from several points of view (e.g. top versus bottom) but could interactively control how detailed a description is either seen or heard regarding that maintenance activity.

Several Interactive Movie videodiscs have been completed using hand-to-hand combat (i.e., karate) as the subject area. These disc let the viewer not only control playing a particular karate move backward and forward at any rate, but also include multiple views and closeup views following every move from four different positions. Several Interactive Movies that focus on equipment maintenance tasks are in progress.

Surrogate Travel (Media Mapping)

Surrogate Travel forms a new approach to local familiarization and low-cost trainers. The basic principle is simple. Up to 108,000 images, showing discontinuous motion along a large number of paths in an area, are stored on a videodisc. Under microprocessor control, the user accesses different sections of the disc, simulating movement over the selected path.

The user sees with photographic realism the area of interest. Unlike a travel movie, the user is able to both choose the path and control the speed of advance through the area using simple controls. The videodisc frames the viewer sees originate as filmed views of what one actually would see in the area. To allow coverage of very large areas, the frames are taken at periodic intervals that may range from every foot inside a building, to every ten feet down a city street, to hundreds of feet in a large open area, e.g., a harbor.

The rate of frame playback, which is the number of times each video is displayed before the next frame is shown, determines the apparent speed of travel. Free choice in what routes may be taken is obtained by filming all possible paths in the area as well as all possible turns through all intersections. While it might first appear that this would be a time consuming and expensive technology, it is in fact relatively efficient because of the design of special equipment and procedures for doing the filming.

Demonstrations of this technology have been developed for building interiors (MIT, National Gallery of Art), a small town (Aspen, Colorado). an industrial facility (nuclear power plant), a weapon site, San Francisco Harbor and the Delta Corridor area of the Marine Corps Air Ground Combat Center (MCAGCC), Twentynine Palms, California.

To provide training in reading and understanding maps, the photograph-based Surrogate Travel, or Media Map, is linked to different sorts of maps of the area. In effect, the viewer can travel across a map, can focus in on it, getting greater and greater detail from what can be presented by standard map symbology, and then "fall through" the map to see photographically what the map depicts. In addition, the viewer can switch among different types of maps (e.g., topographic, infrared, etc.) to develop an understanding of how different map symbologies and representations interact.

In addition to ground level travel, including the inside and outside of buildings, aerial flight experience can be produced and used to simulate aerial reconnaissance or for flight training. Similarly, other forms of travel experience, such as anchorage piloting and low level nap of the earth flying. are also easily accommodated.

Surrogate travel can also be used to provide training on routine and emergency procedures, physical plant maintenance, safety, security as well as other training requirements found in ships, military and industrial facilities. In these applications blueprints, floor plans, procedures, and up-to-date reference materials are linked with the photography of the site to provide a powerful and easy to use training system.

Electronic Libraries

Electronic libraries, in the form of Spatial Data Management Systems (SDMS) provide students and instructors with quick and easy access to an assortment of multi-source and multi-media information. Users literally "fly over" information and select what they want by simply pointing. Spatiality is used to group materials into lesson plans so that different information spaces represent course concept, additional instruction, and assessment procedures.

Stored on a videodisc are tens of thousands of frames consisting of photographs, diagrams, charts, texts, movies, spoken speeches, music, graphs, etc. The pages can be organized, reassembled, segmented, and/or duplicated in accordance with the user's need and growing sophistication with the subject matter. The pages can be annotated, highlighted, drawn-on, underlined, etc., at the user's convenience and pleasure.

For the instructor, the SDMS provides ready access to a wealth of material which might otherwise be unaccessible. Instructors can access the SDMS to create their own information spaces (i.e., courses or lectures) and subsequently present such materials to large audiences in single locations via large screen television projection or to multiple locations through cable distribution systems.

Students can independently use the SDMS for self paced instruction by either working through previously designed information spaces or by browsing on their own. When students and instructors are in remote locations, offsite instruction is facilitated by linking two or more SDMS's together using regular telephone lines. In this manner, a student or instructor can literally fly the other to a topic of interest, sharing at geographically remote sites a large library of information.

The same video materials can be used for hundreds of different users. The only thing that must be changed from user to user is the magnetic storage medium (usually a floppy disc) which serves as the user's private librarian for the videodisc.

MARINE CORPS INTEREST IN VIDEODISC TECHNOLOGY

Marine Corps interest in the potential of videodisc technology stems from a visit by several members of the staff of the Education Center of the Marine Corps Development and Education Command (MCDEC) to the Massachusetts Institute of Technology (MIT) in the Fall of 1978. The Man-Machine Interface Lab at MIT had developed the prototype of the spatial data management system (SDMS) under sponsorship of the Defense Advanced Research Project Agency (DARPA). Education Center personnel sensed the immense potential that interactive videodisc systems held for improving military training and professional education. Initial interest centered around the media mapping and SDMS configurations. The Director, Education Center, subsequently requested that the Marine Corps procure a single small-screen SDMS for the Education Center to use for experimental purposes. The system, which was configured and delivered to the Education Center by Interactive Television Company of Arlington, Virginia, consists of the following equipment:

- Microprocessor-based Spatial Data Management System (microprocessor, keyboard, joystick control, and CRT display).
- MCA Model 7820 Optical Videodisc Player.
- Black and White Video Printer.
- 15" Color Monitor.

It may be worthwhile at this point to briefly review some of the major capabilities of the Marine Corps interactive videodisc system configuration in order to show how these capabilities may be exploited in supporting and enhancing training and education programs:

A. Ease of Operation by User

- No typing or keyboard is required
- User programming is not required
- No query language must be learned
- No training period is required

- Rapid retrieval
- Joystick controls
- Keyboard or symbological data retrieval
- Portable

B. Videodisc

- 54,000 pictures/side
- Each picture/frame has a displayable electronic I.D. number 1-54,000
- Rapid access to any picture/frame (1/8 sec. to 3 seconds)
- Long life, durable, portable
- Thousands of copies of a master is possible
- Dual-speech/sound tracks each side.

C. Applications

- Group instruction with either multiple small screen display and/or large screen display
- Personalized interactive, self-paced learning with built-in testing and tutoring, augmenting or replacing written materials e.g., programmed texts, etc.
- Dual language capability on narrations
- Nonresident instruction via teleconferencing and/or videodisc-based replications of resident instruction
- Lecture/course development - an electronic data base with visual aids which serves as a repository for instructional materials
- Multi-Media - videodisc, magnetic digital disc, and microfiche in combination using text, movies, slides, drawings, animation, graphs, maps, aerial imagery, speech, sound, etc.
- Teleconferencing - phone lines or hardwire
- Management Information System (MIS)/Archival

Repository - can be used as a library of archival materials, for storage and retrieval of operation plans, orders, maps, historical documents, photography, statistical data, important papers, directives, legal records, etc. either classified or unclassified.

- Media Mapping - user controls path, viewpoint, rate of travel, time of day, time of year, weather, mode of travel, detail, and any anciliary information (so long as it has been planned for and included in the program). Media mapping discs can be used for terrain analysis, targeting, simulation, intelligence, anti-terrorist planning, multi-sensor planning and display, tactical situation map/status board, combat in built-up areas, etc.

- Interactive Movies - self-paced or group instruction, lecture support, tactics training, decision making, equipment maintenance and operator training, simulation, etc.

THE TWENTYNINE PALMS VIDEODISC PROJECT

As its initial project, the Education Center undertook to produce a media map of the Delta Corridor area of the Marine Corps Air Ground Combat Center (MCAGCC) at Twentynine Palms, California. The Marine Corps conducts a number of combined arms exercises in the Delta Corridor every year. Because the terrain is of critical concern to units throughout the Marine Corps, it was concluded that the disc would serve as a solid test bed for the Marine Corps wide use of SDMS technology. Interactive Television Company was eventually awarded the contract to film the terrain and to produce and index the media map videodisc. The disc has been produced and indexed. It will be evaluated during the period January - February 1982.

The Twentynine Palms videodisc contains extensive coverage of the Delta Corridor, the Expeditionary Air Field, and the Camp Wilson area of the Combat Center. It includes ground routes and corresponding aerial views from both 200 and 1000 feet. For planning and instructional purposes maps and photography of major Delta Corridor features are included on the videodisc. Following is a description of the videodisc's contents.

Ground Routes. On the videodisc are four-way views (forward, reverse direction, left and right) of over 100 miles of ground travel.

The ground routes were filmed from a Marine Corps M880 truck with a 30 foot picture interval. The truck was driven at approximately five miles per hour over terrain varying from smooth asphalt roads to very rugged open terrain.

Because of the rugged terrain a gyro stabilizer camera mount was used to provide side and front view images with a constantly even horizon. To date, most viewers have not been able to distinguish between the stabilized and unstabilized views on any of the terrain - paved or rough.

Aerial Routes. A marine Corps HU-1 helicopter was used to film corresponding aerial views from the ground routes. Aerial views at 200 and 1000 feet were mapped. Picture intervals of 50 and 250 feet respectively were used for the aerial mapping.

For the aerial photography, a forward-looking camera was mounted to one of the helicopter skids. A side-looking camera was clamped down looking out from the door of the helicopter. An intervalometer was used to control the camera framing rates.

Ground and Aerial Features. Throughout the Delta Corridor selected terrain and tactical features were identified and filmed. Terrain features included examples of saddles, crests (military and topographic), dry washes, passes, contours and so on. Tactical features included panoramic and close-up photography of major objective areas, staging and assault areas, major passes, a firing range, and other areas used in Delta Corridor exercises. Ground and aerial photography of both terrain and tactical features was included on the videodisc.

In the Delta Corridor, the helicopter was used to film panoramic views as would be seen by an observer located on any one of the three Observation Posts overlooking the Delta Corridor area. Observation Post views look outward in a 360 degree sweep. Inward views are also included so users can readily identify the observation posts and know what is located on them.

Ground and aerial photography of the Expeditionary Air Field and Camp Wilson area, which is used as the base camp for units visiting the Combat Center to participate in exercises, are also on the videodisc. Included are views of the runway, support facilities, fuel bunkers, mess and shower facilities, and housing areas.

Maps. Several topographic maps of different scales were filmed and included on the videodisc. These maps can be used to "fly over" the mapped areas moving in and out for greater or less detail. Under microcomputer control, the maps can be correlated to aerial and ground photography.

Narrated Introduction. A narrated introduction at the start of the videodisc describes the project's applications to Marine Corps problems. The introduction also summarizes the videodisc's content.

It is anticipated that the Twentynine Palms Videodisc Media Map will serve multiple training functions within the Marine Corps. For the Marine officer planning or engaged in exercises within the Delta Corridor, the Videodisc Map will serve as an aid to planning and terrain familiarization. For Marine Corps elements (such as the Education Center at Quantico, Virginia) engaged in teaching land navigation, terrain analysis, and map reading, the Videodisc Map will provide the instructor with a valuable training aid that integrates ground, aerial, and map materials for instruction.

FUTURE MARINE CORPS PROJECTS

The Marine Corps could select any one of a number of useful videodisc applications for its next project. The following is a partial list of possible applications:

. Surrogate travel through a Landing Helicopter Assault ship (LHA) to aid in the pre-embarkation orientation of selected spaces and operational features of interest to the Landing Force. Similar discs could be developed for other appropriate types of amphibious ships.

. Interactive movie photography, incorporating surrogate travel techniques, which would enable students to become familiar with high-cost, low-density equipment that they ordinarily would not get an opportunity to work with in the classroom.

. Interactive videodisc-based similations of selected USMC command and control (shore and seabased) facilities.

. An electronic library of resource information collected for threat analysis purposes. The disc-based information could supplement or replace existing hard copy media for instructional development, self- paced student research, reference or instruction.

. An electronic library of information, including media map material, of terrain, lines of communications, port facilities, beaches, landing zones, airfields, and cities to be combined with other maps, photography and target data required to support amphibious operation planners, instructors, and students involved in war gaming exercises.

. An electronic library of Marine Corps archival material relating to the evolution of amphibious operations. This would serve a two-fold purpose in that it would: 1. ensure against the loss of critical information by providing working copies of historical documents; 2. establish a unique cross reference and information retrieval system for researchers and amphibious planners.

It should be recognized that the videodisc is a storage medium, not an instructional system. But it is a storage medium of such remarkable capacity and with such varied capabilities that, if used with imagination in support of carefully defined objectives, it can elevate existing instructional systems to new levels of excellence. As the Marine Corps continues to define its training objectives, it undoubtedly will identify new applications for the videodisc and related forms of telecommunication technology. In the meantime, it will continue to monitor the exploratory efforts of the Army, the Air Force, and other agencies in and out of government that are developing new uses for this powerful medium.

TESTING DURING TRAINING:
WHY DOES IT ENHANCE MOVEMENT RETENTION

Joseph D. Hagman
US Army Research Institute for the Behavioral and Social Sciences

Linear movement retention was examined for training methods emphasizing (repeating) either presentation (p) or test (t) trials. P-trials were experimenter-defined study movements constrained by a mechanical stop; t-trials were learner-defined recall movements unconstrained by the stop. Separate groups of governmental employees received training consisting of three, 6-trial cycles. Cycles began with a p-trial that defined the criterion movement to be remembered. The five remaining trials of each cycle varied in type across groups. One group, for example, performed successive t-trials, whereas another performed successive p-trials yoked in value to the first group's t-trials. Retention was then examined at 3 minutes and 24 hours after training. Absolute (unsigned) error revealed that t-trials were more effective than yoked p-trials in promoting movement retention. The data were consistent with the hypothesis that retention benefits obtained from testing during training result from better initial learning (encoding) of kinesthetic cues generated under a learner-defined than under an experimenter-defined movement execution mode. It was concluded that testing cannot only be used to evaluate but also to improve motor skill retention.

TESTING DURING TRAINING: WHY DOES IT ENHANCE MOVEMENT RETENTION

JOSEPH D. HAGMAN
US ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
ALEXANDRIA, VA 22333

The Army's primary peacetime mission is to maintain combat readiness (Guthrie, 1979). To be combat ready, soldiers must first become proficient in their performance of job tasks, and then, retain this proficiency over what can be prolonged periods of no practice. One way to enable soldiers both to reach and maintain combat readiness is through the use of task training methods that promote effective acquisition and retention. To do this, these methods must be identified and compared.

A review of the training research literature reveals that training methods have been compared primarily within the context of laboratory experiments. Here, training has involved the execution of presentation (p) trials, where to-be-learned information is presented by the experimenter to the learner for study, and test (t) trials, where this information is removed and the learner attempts to recall (reproduce) it from memory. Although standard training methods involve alternation of p- and t-trials (e.g., Tulving, 1967; Wrisberg & Schmidt, 1975), the most effective number and sequential arrangement of p-and t-trials to use is a matter of debate. From a traditional learning theory viewpoint, where p-trials are seen as having an effect similar to reinforcement (Adams & Dijkstra, 1966), training methods that emphasize (repeat) p-trials should be more effective than those that repeat t-trials. P-trial repetition increases the number of reinforcement opportunities during training, and therefore, should enhance both acquisition and retention. From a contemporary cognitive learning viewpoint, on the other hand, information processing activities such as memory retrieval and internal item generation are considered important aspects of acquisition and retention (Bjork, 1975). Because t-trials provide an opportunity to perform these activities on information studied during p-trials, training methods that repeat t-trials should also be effective.

P-trial effects have been documented in numerous experiments showing that improved performance occurs when p-trials are repeated during training (e.g., Adams & Dijkstra, 1966). Only recently, however, have improvements associated with t-trial repetition been reported. Researchers have shown that with verbal tasks t-trials not only contribute to acquisition (e.g., Lachman & Laughery, 1968) but also to retention (Hogan & Kintsch, 1971; Wenger, Thompson, & Bartling, 1980). Even more recently, t-trials have been reported to influence motor task performance. Hagman (1980a,b), for example, had persons learn either the distance (extent) or end-location (terminal position) of linear positioning movements under training methods emphasizing either p- or t-trial repetition. P-trials were movements terminated by a mechanical stop that was prepositioned by the experimenter to ensure execution of the to-be-learned criterion movement cue (i.e., distance or end-location). T-trials were movements performed with the stop removed. It was during t-trials that learners stopped their own movement when they thought they had accurately recalled the criterion movement cue. Results of both experiments showed that movement cue acquisition was better when p-trials were repeated during training, whereas long-term retention was better when t-trials were repeated during training.

The purpose of the present experiment was to extend these earlier findings by testing two hypotheses suggested (Hagman, 1980) to account for the beneficial effect of t-trials on movement cue retention. The first hypothesis relies on the procedural distinction between experimenter-defined (i.e., performed with the stop present) and learner-defined (i.e., performed with the stop absent) movements. Evidence suggests that movement cues generated under a learner-defined execution mode are retained better than those generated under an experimenter-defined execution mode (Kelso, 1977; Stelmach, Kelso & McCullagh, 1976). This enhanced retention is caused by superior learning (encoding) of learner-defined movement cues brought about by the learner's ability to predict or anticipate cue values prior to movement initiation (e.g., Kelso, 1977). T-trials allow for prediction because they are learner-defined, whereas p-trials do not allow for prediction because they are experimenter-defined. In a multitrial training context learners base posttraining recall attempts on their retention of cues generated during the trial type repeated during training. That is, learners rely on p-trial retention when p-trials are repeated, whereas they rely on t-trial retention when t-trials are repeated. Because t-trials are learner-defined, retention of t-trial generated cues should be superior to retention of p-trial generated cues which are experimenter-defined. Thus, enhanced long-term motor retention should occur with training methods that emphasize learner-defined t-trial repetition.

The second hypothesis proposed to account for the beneficial effect of t-trial repetition on movement cue retention involves the notions of movement variability and motor schema. The motor schema is an abstraction of task and environmental characteristics that develops through repeated and varied movement during training (Schmidt, 1975), and serves as a rule or concept for movement generation. Researchers have found that as variability increases during training the abstracted schema information becomes increasingly resistant to forgetting (Newell & Shapiro, 1976; Posner & Reele, 1970). In the previous experiments by Hagman (1980a,b), variability during training was generated at t-trials because learners were inconsistent in their recall attempts. In contrast, no variability was generated by p-trials because all were identical in terms of distance (Hagman, 1980a) or end-location (Hagman, 1980). As a result, it could be argued that schema strength was greater after repeated t-trial training than after repeated p-trial training. Thus, one would predict better retention under the former than under the latter training method.

The general approach used in the present experiment to test the validity of these two hypotheses involved yoking separate p-trial training method groups to both the t-trial distance and t-trial end-location groups trained earlier. Yoking involved using a mechanical stop to ensure that p-trials of the yoked groups were identical to the t-trials of the other groups in terms of both distance and end-location. Thus, yoking afforded the means of equating p- and t-trials in terms of variability during training but allowed the distinction to remain between p- and t-trial execution mode (i.e., experimenter- versus learner-defined). If variability per se during training is the key to enhanced retention of movement cues, then one would expect the retention displayed by the two yoked p-trial groups not to differ from that displayed by the two t-trial groups. If, on the other hand, movement execution mode during training is the key to enhanced retention, then one would expect the two t-trial groups to display retention superior to that of the two yoked p-trial groups.

Method

Subjects

Sixty governmental employees volunteered to serve as participants in the experiment. All were members of the professional and clerical staff of the Army Research Institute for the Behavioral and Social Sciences.

Apparatus

Participants were required to make movements from left to right using a metal slide that ran along a linear track consisting of two stainless steel rods 35 inches (88.9 cm) in length. Two Thompson Ball Bushings supported the slide on the rods which were mounted in parallel on a metal frame 4.25 inches (11 cm) apart and 11 inches (27.94 cm) above the frame base. The base rested on a standard table top 31 inches (78.74 cm) from the floor. A second slide was used by the experimenter to stop movement of the first slide along the track. A pointer attached to the experimenter's side of each slide ran along a meter stick to indicate respective slide position. Additional apparatus included a chin rest to stabilize head position, earphones through which tape-recorded procedural commands were delivered, and a blindfold to eliminate visual cues.

Design

The experiment contained an acquisition and a retention segment as shown in Figure 1. The acquisition segment consisted of 18 training trials divided into three cycles of six trials each. Cycles contained p- and t-trials. P-trials were experimenter-defined movements terminated by the mechanical stop. The stop was prepositioned by the experimenter to ensure that participants executed (studied) the criterion distance end-location at p-trials and duplicated t-trials at yoked p-trials, i.e., py. T-trials were learner-defined recall movements unconstrained by the mechanical stop. Four training method groups were included in the experiment, i.e., DISTANCE PRESENTATION (DP), DISTANCE TEST (DT), END-LOCATION PRESENTATION (LP), and END-LOCATION TEST (LT). Training methods differed in their emphasis on p- and t-trials performed during each cycle. Group DT performed cycles containing an initial to-be-learned criterion p-trial followed by five successive recall t-trials. Group DP performed cycles containing six successive p-trials. The first was the criterion, but the next five were yoked in distance to the corresponding t-trials of Group DT. Yoking was also applied to the two end-location groups in a similar fashion. Because of this yoking

	ACQUISITION			RETENTION	
GROUPS	CYCLE 1	CYCLE 2	CYCLE 3	3 MINUTES	24 HOURS
DP	P P_Y P_Y P_Y P_Y P_Y	P P_Y P_Y P_Y P_Y P_Y	P P_Y P_Y P_Y P_Y P_Y	T	T
DT	P T T T T T	P T T T T T	P T T T T T	T	T
LP	P P_Y P_Y P_Y P_Y P_Y	P P_Y P_Y P_Y P_Y P_Y	P P_Y P_Y P_Y P_Y P_Y	T	T
LT	P T T T T T	P T T T T T	P T T T T T	T	T

*P_Y=YOKED

Figure 1. Trial sequence for training method groups at acquisition and retention

procedure, Groups DT and LT were trained before Groups DP and LP. Data from the two yoked PRESENTATION groups were collected in the present experiment, whereas data from the two TEST groups were collected earlier (Hagman, 1980a,b). Although trained at different times, subjects in the two yoked groups were drawn from the same population as those in the two TEST groups.

The retention segment of the experiment consisted of a single t-trial performed by each group at both 3 minutes and 24 hours after acquisition, as shown in Figure 1. Separate 2x2 mixed factorial designs were used to examine distance and end-location cue retention. The between-subjects factor was group (DP, DT, or LP, LT) and the within-subjects factor was retention interval (3 minutes, 24 hours). Fifteen participants were assigned to each of the four training method groups with the constraint that each group contain the same proportion of men and women.

Procedure

Participants were instructed to learn and remember either movement distance or end-location depending on their group. Those in groups DP and LP were also told of the yoking procedure. All participants were then shown a written copy of the trial command sequence that they would be hearing and told the meaning of each command. The p-trial command was "Movement" and the t-trial command was "Recall Movement." Each of these commands was preceded by "Ready" and followed by "Rest." At "Ready" the experimenter grasped the participant's hand and placed it on the handle of the slide. Five seconds later, the participant heard either "Movement" or "Recall Movement" depending on the trial type. At "Movement," participants moved the slide across the track until contacting the mechanical stop. At "Recall Movement," those in Groups DT and LT moved the slide across until they felt that they had recalled the criterion distance or end-location, whereas those in Groups DP and LP moved the slide along until contacting a stop. This stop was prepositioned by the experimenter at the distance or end-location recalled by participants in Groups DT and LT at t-trial execution. Five seconds were allowed for movement execution. During this interval, participants received white noise through earphones to eliminate auditory cues resulting from displacement of the slide. "Rest" marked the beginning of a 10-second interval during which participants removed their hand from the slide and placed it on the table in a predetermined resting position. During rest periods the experimenter recorded recall accuracy to the nearest millimeter (when appropriate) and repositioned the stop in preparation for the next trial. After "Rest," participants heard "Ready" and the command sequence for the next trial began. During the retention segment of the experiment, intervals of 3 minutes and 24 hours were inserted between "Rest" and "Ready." In general, participants were instructed not to count during movements and shown the approximate movement speed (i.e., 125 mm/sec) desired by the experimenter. Prior to making the first movement, participants donned their blindfold and earphones, and then were given a 10-second opportunity to move the slide and get a feel for its movement characteristics.

Results

Algebraic (signed) and absolute (unsigned) error scores were recorded for each t-trial performed during the retention segment of the experiment. No acquisition data were analyzed because yoking prevented any differences in group performances. Each performance measure was analyzed separately.

Retention was examined using a 2x2 mixed factorial Group (DP, DT or LP, LT) by Retention Interval (3 minutes, 24 hours) analysis of variance (ANOVA). Separate ANOVAs were performed on the algebraic and absolute error scores for the two distance groups (DP, DT) and the two end-location groups (LP, LT). No significant ($P < .05$) effects of interest were found for algebraic error, therefore only absolute error scores are reported.

Distance. Mean absolute error scores are shown in Figure 2. The scores for the two distance groups (i.e., DP, DT) are on the left and those for the two end-location groups (i.e., LP, LT) are on the right. The absolute error ANOVA revealed no significant main effects for a significant groups x retention interval interaction, $F(1,28)=6.85$. The rejection region for this and all other analyses was .05. This interaction resulted from an increase in recall error over time for Group DP and an associated decrease in recall error over time for Group DT. Individual comparisons of simple main effects using the least significant difference method (Carmer & Swanson, 1973) revealed that the Group DP error increase was revealed that 3 minutes after training no difference in recall error existed between Groups DP and DT, whereas 24 hours after training Group DP displayed greater recall error than that of Group DT.

End-location. The absolute error ANOVA, for end-location, revealed a significant main effect of group, $F(1,28)=5.85$, demonstrating greater posttraining recall error for Group LP than for Group LT, and a group x retention interval interaction that approached significance, $F(1,28)=3.11$, $.05 < p < .10$. Although nonsignificant by conventional standards, further analysis of simple main effects associated with this interaction was justified by a priori expectations of training method outcome as indicated by the results obtained for distance cue recall. As shown in Figure 2, the marginal interaction resulted from an increase in recall error after training for Group LP while Group LT error remained almost unchanged. Individual comparisons revealed that the Group LP increase was significant, and that Group LT error was statistically stable. Group recall performance did not differ 3 minutes after training while 24 hours after training Group LP error was significantly greater than Group LT error. Conservatively speaking, the absolute error data for both movement distance and end-location cues reveal that training methods that emphasize testing (i.e., DT, LT) prevent posttraining task retention decrements, whereas those that emphasize presentation produce marked posttraining retention decrements. Thus, even the yoking procedure used in the present experiment to increase movement variability during training was unable to prevent forgetting when p-trials were emphasized.

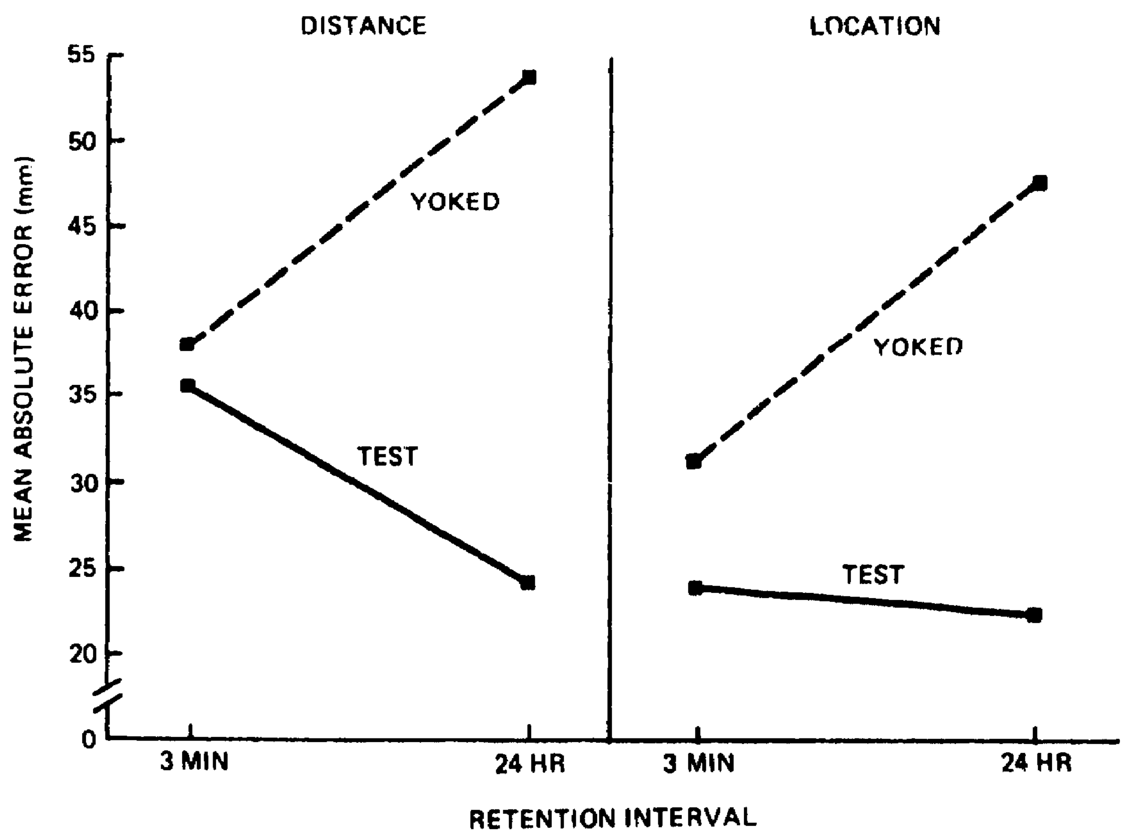


Figure 2. Mean absolute error on retention t-trials for distance and end-location training method groups.

Discussion

The purpose of this experiment was to explain previous data showing that repeated testing during training is more effective than repeated presentation in promoting long-term motor task retention (Hagman, 1980a,b). Two hypotheses were tested. The first stated that retention benefits were caused by differences in the learning (encoding) characteristics of p- and t-trial due to differences in movement execution mode. The second hypothesis stated that retention benefits were the result of increased movement variability produced by t-trial execution during training. The present absolute error differences found between Groups DP and DT and between Groups LP and LT support the execution mode hypothesis. Although p- and t-trial variability was equated during training through yoking, retention differences at 24 hours after training still favored the t-trial repetition groups for both distance and end-location cue recall. Thus, the variability hypothesis is not supported.

How does movement mode influence retention? As suggested earlier (Hagman, 1980b), in multitrial training situations where either p- or t-trials are emphasized through repetition, learners base later recall attempts on their retention of movement cues generated at repeated trials. It is easier to remember t-trial cues than p-trial cues because the former are learner-defined. Better retention of learner-defined cues comes from the learner's ability to predict or anticipate movement cues prior to initiation. According to Kelso (1977), "when a person is able to predict movement, two sets of signals are generated; (a) the downward discharge to effector organs, and (b) a simultaneous central discharge from motor to sensory centers that presets sensory systems for the anticipated consequences of the motor act" (p.35). Thus, the role of anticipation or prediction is to enhance the encoding of movement kinesthetic information arising from muscles and joints (Kelso, 1977; Stelmach, et. al., 1976). An extension of this corollary discharge theory can explain the superior retention resulting from t-trial repetition. It is argued that at t-trials cortical sensory centers are more prepared to receive incoming afferent impulses from muscles and joints, since movement consequences can be anticipated. At p-trials, on the other hand, this would be more difficult since little if any prior information is available regarding the terminal locus of the movement. It is this superior encoding of t-trial cues relative to p-trial cues that causes superior long-term retention.

Finally, it should be mentioned that although the present results rule out variability per se as the cause of t-trial retention effects, they do not rule out the possibility that variability contributes to retention, but does so only when generated during learner-defined movements. It could be argued, for example, that the effects of variability are dependent on movement mode, and perhaps vice versa. Although the present experiment does not discount this interpretation, no data have been reported either to suggest or support it. Therefore, it remains highly speculative, yet worthy of future research.

Conclusions

The results of this experiment help to clarify past research findings and answer the question of why testing during training enhances motor task retention. In doing so, they assist the Army in its quest to identify training methods that produce the highest levels of motor task acquisition and retention.

From the results it can be concluded that: (a) Training methods that provide for increased opportunities for testing improve long-term motor task retention; (b) these benefits derive from the superior encoding of learner-defined movements performed during t-trials, relative to experimenter-defined movements performed during p-trials; (c) increased variability of movement caused by t-trial repetition during training is not responsible for the obtained retention benefits associated with testing; (d) testing during training benefits both movement distance and end-location cue retention.

References

Adams, J. A., & Dijkstra, S. Short-term memory for motor responses. Journal of Experimental Psychology, 1966, 71, 314-318.

Bjork, R. A. Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), Information processing and cognition the Loyola symposium. Hillsdale, N. J.: Erlbaum, 1975.

Carmer, S. G., & Swanson, M. R. An evaluation of ten pair-wise multiple comparison procedures by Monte Carlo methods. Journal of the American Statistical Association, 1973, 68, 66-74.

Guthrie, J. R. Supporting the Army in the eighties. Army Logistician, September-October 1979, 11(5), 22-25.

Hagman, J. D. Effects of Presentation- and Test-Trial Training on Motor Acquisiton and Retention. (Technical Report 431) Alexandria, Va.: US Army Research Institute for the Behavioral and Social Sciences, January, 1980(a).

Hagman, J. D. Effects of Presentation- and Test-Trial Training on Acquisition and Retention of Movement End-Location. (Technical Report 492) Alexandria, Va.: US Army Research Institute for the Behavioral and Social Sciences, November, 1980(b).

Hogan, R. M. & Kintsch, W. Differential effects of study and test trials on long-term recognition and recall. Journal of Verbal Learning and Verbal Behavior, 1971, 10, 562-567.

Kelso, J.A.S. Planning and efferent components in the coding of movement. Journal of Motor Behavior, 1977, 9, 33-47.

Lachman, R., & Laughery, K. R. Is a test trial a training trial in free recall learning? Journal of Experimental Psychology, 1968, 76, 40-50.

Newell, K. M., & Shapiro, D. C. Variability of practice and transfer of training: Some evidence toward a schema view of motor learning. Journal of Motor Behavior, 1976, 8, 233-243.

Posner, M. I., & Keele, S. W. Retention of abstract ideas. Journal of Experimental Psychology, 1970, 83, 304-308.

Stelmach, G. E., Kelso, J. A. S., & McCullagh, P. D. Preselection and response biasing in short-term motor memory, Memory and Cognition, 1976, 4, 62-66.

Tulving, E. The effects of presentation and recall of material in free recall learning. Journal of Verbal Learning and Verbal Behavior, 1967, 6, 175-184.

Wenger, S. K., Thompson, C. P., & Bartling, C. A. Recall facilitates subsequent recognition. Journal of Experimental Psychology: Human Learning and Memory, 1980, 6, 135-144.

Wrisberg, C. A., & Schmidt, R. A. A note on motor learning without post-response knowledge of results. Journal of Motor Behavior, 1975, 7, 221-225.

Hall, Eugene R. & Freda, Jon S., US Navy Training Analysis and Evaluation Group, Orlando, Florida. (Tues. A.M.)

Differential Effectiveness and Efficiency of Individualized Instruction: I. Background and Study Design

The Navy's Training Analysis and Evaluation Group (TAEG) conducted an empirical study to determine if measures of training effectiveness or training efficiency were differentially related to the method of instruction used to train Navy technical school graduates. Also investigated were the interrelationships between methods of instruction, ability levels of graduates and the types of tasks taught in the technical schools. Since the same content was not taught in courses conducted under each method, it was necessary to equate the content of the 20 courses (i.e., tasks taught) to a common base so that appropriate comparisons could be made. For this purpose a generic task classification system was developed. School achievement and job performance measures obtained on over 5000 technical school graduates served as the criterion variables. This paper presents the background for and design of the study.

DIFFERENTIAL EFFECTIVENESS AND EFFICIENCY OF INDIVIDUALIZED INSTRUCTION: I. BACKGROUND AND STUDY DESIGN

Eugene R. Hall and Jon S. Freda

Training Analysis and Evaluation Group
Orlando, Florida 32813

Individualized instruction has become a controversial issue in military training. Many individuals in both training and operational settings have come to believe that individualized instruction (II) is not a desirable or effective way to train students for operational job assignments. The widespread belief is that conventional classroom group-paced (GP) methods result in better trained personnel.

Currently, the U.S. Navy conducts technical training under both II and GP instructional methods. The principal II methods used are computer managed instruction (CMI) and self-paced (SP) instruction. Because of the potential for reduced student training time, the Navy plans to individualize still more of its courses. However, in view of concerns expressed by fleet units, the Chief of Naval Education and Training (CNET) tasked the Training Analysis and Evaluation Group (TAEG) to conduct a study to examine the effects of individualized instruction.

Purpose

The purpose of the study was to determine if individualized instruction is more or less effective and/or efficient than conventional instruction, and further, if these effects differentially relate to:

- training individuals of differing ability levels and/or
- training different types of tasks.

The present paper presents details of the methodology employed to conduct the study. A subsequent paper in this volume (Freda, Hall and Ford) presents and discusses study findings.

Method

Since experimental methods were impractical to employ within the context of an ongoing military training system a correlational approach was used to conduct the study. Under this approach, statistical analyses are performed on existing record data to determine if significant relationships exist among variables of interest. Because of limitations inherent in available data however, caution is required in the interpretation of results obtained. Also, there are limits on the generalizations that can be made. While some results may be viewed as definitive, others must be viewed as only suggestive. Further, certain of the results should be considered applicable to only the courses included in the study. These situations will be appropriately noted in the present text and in the subsequent paper by Freda, Hall, and Ford.

Study Questions

The TAEG study was designed to answer two major questions derived from the orginal CNET tasking:

1. Are significant, relative differences in training effectiveness and/or efficiency associated with individualized instruction versus group-paced (GP) instruction?

2. If there are differences, are they related to different kinds of training tasks and/or ability levels of trainees?

Study Variables

The study examined seven major variables. The major predictor variables were method of instruction, trainee ability level, and type training task. The major criterion variables were training costs, end of course grades, time to complete the course, and Naval Education and Training Command (NAVEDTRACOM) Training Appraisal System (TAS) ratings (see below). Training costs and time to complete were viewed as training efficiency measures, and end of course grades and TAS ratings as internal and external measures of training effectiveness, respectively.

Predictor Variables.

Method of Instruction. The primary predictor variable of the study was instructional method. Two basic methods were of interest: Individualized Instruction (II) and Conventional Instruction (CI). II is defined as an instructional strategy in which the learning activities are designed to accommodate individual differences in background, skill level, aptitudes, and cognitive styles. For the present study, II involved self-paced and computer-managed courses. CI is defined as an instructional strategy in which learning activities are directed toward a normative model of the target population characteristics. CI usually occurs in a group environment (i.e., is group-paced).

To develop lines of inquiry for the study, a brief literature survey was conducted. Previous findings from reviews of II versus CI used in military settings (e.g., Orlansky & String 1979) indicate that:

- II is as effective as CI in terms of end of course achievement scores
- the efficiency of II is significantly better than that of CI in terms of student time to complete instruction
- training costs for II are lower due to the student time savings.
- the addition of computer support (either CAI or CMI) to self-paced instruction does not significantly increase student time savings.

Orlansky and String also noted that the previous research contrasting II and CI was confined to comparisons on measures of student learning achievement available at the schools. External criteria of training effectiveness were not used for assessing the effects of different instructional methods.

The relevance of previous research for the TAEG study then was that:

- school performance scores (i.e., end of course grades) should be equivalent for CI versus II graduates overall
- time to complete training should be significantly less for II than for CI
- student time to complete training should be equivalent for SP versus CMI.

Since no previous data were reported for method effects on external criteria of training effectiveness, these analyses were considered wholly exploratory.

Ability Level. Ability levels of trainees constituted the second major predictor variable of interest. For the present study, ability levels of trainees were considered to be reflected by student scores on the Armed Services Vocational Aptitude Test Battery (ASVAB). This test is routinely administered to all armed services enlistees.

The ASVAB consists of 12 subtests. Various combinations, or composites, of ASVAB subtest scores are used by the Navy to determine an individual's eligibility for attendance at specific technical schools. Scores from three of the battery subtests can also be used to derive an Armed Forces Qualification Test (AFQT) percentile equivalent. These percentile equivalent scores provide a measure of general ability. They can also be used to group individuals into mental categories. AFQT percentile scores (derived from the pre-October 1980 conversion routines) were used to represent ability levels of graduates for making comparisons across courses.

Types of Training Tasks. Definitive conclusions concerning the value of II versus CI for training different types of tasks require that the same tasks be taught under each method. Unfortunately, the same courses (and, consequently, the same tasks) are not taught under each of the two basic methods of instruction, and it was beyond the scope of the study to create this condition experimentally. Thus, the matching of tasks under different methods of instruction was approximated through the use of a generic task classification system. The content of the different courses (i.e., tasks taught) was equated to a common base so that appropriate comparisons could be made. The use of these comparisons assumes that the psychological processes involved in the acquisition of generic skills (e.g., decision making) are the same regardless of the specific context in which the learning occurs.

For the present TAEG study, a modification of the Instructional Quality Inventory (IQI) method (Ellis, Wulfeck and Fredericks, 1979) was used to equate skill and knowledge items taught in each of the courses to a common base. Subject matter experts (SME) at each school classified statements of the schools learning objectives into one of five types of information content: fact, category, procedure, rule and principle. The IQI definitions of content type (see Ellis, et al) were retained. The five classifications were then interpreted as generic tasks so that comparisons could be made across the courses.

Criterion Variables.

Training Effectiveness Measures. Two sets of measures of students' learning achievements, which indicate course training effectiveness were of interest to the study:

- internal measures which reflected how well students performed in school

- external measures which reflected graduate job performance

Internal measures of student learning achievement were obtained from the schools. These consisted principally of end of course grades. Other internal measures of course data reflecting training effectiveness were not routinely available. External measures of student learning achievement were obtained from the NAVEDTRACOM Training Appraisal System Data Base. These consisted of fleet supervisor ratings of the adequacy of school training for particular tasks which graduates are expected to perform on the job.

The CNET Special Assistant for Training Appraisal (CNET 015) routinely collects feedback data via mailout questionnaires from first-level fleet supervisors of recent (i.e., 4 to 10 months after graduation) technical school graduates. Random samples of graduates are drawn from the total pool of course graduates during a given time frame. Fleet supervisors rate on a 5-point scale (a "1" equals "unsatisfactory" and a "5" equals "much more than satisfactory") the adequacy of school training for an identified course graduate. Training adequacy judgments are made for a number of specific tasks for which a given technical school provided training. The task statements listed on a feedback questionnaire are currently prepared by technical training staff for a given course. The statements are based on the learning objectives of that course.

The TAS data base contained both a listing of tasks taught at a school, and ratings of training adequacy for these tasks. The questionnaire task statements which reflect specific skills and knowledges taught at a school were classified by school SMEs into the generic task categories described above. Thus, specific generic tasks could also be matched to supervisor training adequacy ratings on specific items to permit comparisons within and between courses.

Training Efficiency Measures. Two measures were used to reflect training efficiency: training costs and student time (contact hours of instruction) to complete a course. Training cost data were obtained from the CNET "Per Capita Cost Data Base" for each course in the study. The training cost data were referenced to fiscal year 1979 which coincided with the time period that graduates of interest to the study were in training. From the cost data base, a "training cost to produce one graduate per course session" was derived and used in subsequent analyses. Student course completion times were obtained for individualized courses from CMI files and from class records at the schools for SP courses. Completion times for CI courses were obtained from the schools and from CNET computer files.

Table 1 identifies the major predictor and criterion variables of the study. The measures of the variables are listed as are the data sources. More complete descriptions of the study samples are presented below.

TABLE 1. MAJOR PREDICTOR AND CRITERION VARIABLES

VARIABLE	MEASURES	SOURCE
PREDICTOR		
Ability Level	AFQT, Composites	CNET 015
Method of Instruction	SP, CMI, GP	CNET, CNTECHTRA
Type of Training Task	Five Generic Tasks	IQI (SMEs)
CRITERION		
EFFECTIVENESS		
End of Course Grades	Final Grades	School Records (SP+GP), CNTECHTRA (CMI)
Training Adequacy Ratings	1-5 Scale	TAS (CNET)
EFFICIENCY		
Cost of Course	Average Cost to Produce 1 Graduate (per session)	CNET Accounting System
Time to Complete	Contact Hours	CNTECHTRA (CMI): School Records, SMEs, NITRAS

Study Samples

Course Samples.

The TAEG project staff coordinated with CNET and Chief of Naval Technical Training (CNTECHTRA) staff to select courses for inclusion in the study. The plan was to identify approximately 10 A-level courses that used conventional (i.e., group-paced) instruction and 10 others that used individualized instruction (i.e., self-paced or computer-managed). Courses in each instructional category were to include the full range of ability levels of individuals who undergo Navy technical training. It was also desired that courses in each category be roughly matched on general instructional content (i.e., type training tasks) and on geographic location.

Initial selection of courses for inclusion in the methods of instruction groups was made on the basis of entries in the Navy Integrated Training Resources and Administrative System (NITRAS). This system identifies courses that are considered to be individualized (SP or CMI) and those considered to be taught conventionally (i.e., GP). It should be noted that these instructional methods are not purely applied to all portions of the courses.

Courses selected for the study were those classified as "A-level." "A-level" courses provide basic skill and knowledge training for entry level Navy jobs. These courses were studied rather than C-level courses because:

- a greater diversity of tasks is trained in the more general A-level courses than in the largely equipment-specific C-level courses

- a wider range of student abilities is involved in A-level taining

- proportionately more A-level courses are taught under individualized instruction

It was further desired that courses be selected to the extent possible from those for which TAS data were already available or soon to be available (i.e., within approximately the next 6 months). Although CNET 015 training appraisal schedules could be altered, this would result in lengthy time delays to obtain data on course graduates and would disrupt the ongoing work of the schools and CNTECHTRA staff codes.

Twenty-three A-level courses were nominated by CNTECHTRA for inclusion in the study. Although availability of the TAS data was the major determinant of the sample composition, it is believed that the other criteria were reaonably well met and that the courses constitute a fair sample (approximately 15 percent) of Navy "A" schools. Two officer courses were also selected for study. These courses, "Damage Control Assistant", were ostensibly the same course taught under a different method of instruction at each of two locations. In addition, data were obtained from two basic CMI courses. Graduates of the RM Sea "A" School and the RM Shore "A" School also attended the RM Basics course. Their records were obtained from CNTECHTRA CMI files and used in analyses of interest to the study. Similarly, graduates of the EN, MM 600 psi and MM 1200 psi schools attended the Propulsion Engineering (PE) Basics course prior to entry into their respective "A" schools. Their records were also obtained from CMI files.

Graduate Samples

Information pertaining to the number of graduates per course for whom data were available for selected variables is presented in table 2. Data were obtained from a total of 7,083 records of enlisted personnel and officers who were graduated from the schools between August 1978 and April 1980. This total included duplicate records of graduates from basic courses and the dual-phase HT courses.

TABLE 2. NO. OF GRADUATES PER COURSE FOR WHOM DATA WERE AVAILABLE FOR SELECTED VARIABLES

Course	CDP	Type[1]	Graduating Dates	Original Sample Size	ASVAB	Time to Complete[3]	End of Course Grade	Ratings on TAS Items	Comments
AK	6522	SP	10/79-4/80	254	241	231	0	170	No end of Course Grade (EOCG)
DCA (Phil)	3218	SP	10/79-3/80	60	0	59	54	43	Officer Course No AFQT
DK	6061	SP	8/79-2/80	110	107	110	110	79	
HT2 (Phil)	6339	SP	10/78-3/79	319	318	294	303	218	Same graduates in CDP 6120
IM	6046	SP	10/79-4/80	18	18	17	16	15	
MM 600	6493	SP	10/78-3/79	374	372	284	282	138	PE Basics-MM (CDP 6262) EOCG included in the EOCG for CDP 6493
MM 1200	6492	SP[2]	9/78-3/79	204	203	176	176	124	Same as above for CDP 6492
PN	6102	SP	11/78-5/79	84	74	69	0	65	No EOCG
TD	6521	SP	1/79-7/79	185	179	185	0	162	No EOCG
YN	6057	SP	11/78-5/79	222	208	211	0	174	No EOCG
AD	6591	CMI	10/79-4/80	464	450	454	464	98	
EN	6487	CMI	3/79-6/79	359	356	349	244	192	
HT2 (San Diego)	6106	Mixed	10/78-4/79	264	260	261	190	171	Same graduates in CDP 6119
AE	6515	GP	3/79-5/79	90	86	90	90	54	
ASE	6530	GP	8/79-2/80	36	29	36	36	29	

TABLE 2. NO. OF GRADUATES PER COURSE FOR WHOM DATA WERE AVAILABLE FOR FOUR SELECTED VARIABLES (con't)

Course	CDP	Type[1]	Graduating Dates	Original Sample Size	Selected Variables: ASVAB	Selected Variables: Time to Complete[3]	Selected Variables: End of Course Grade	Selected Variables: Ratings on TAS Items	Comments
AW	6537	GP	1/79-7/79	47	42	45	45	35	
DCA (Treasure Island)	3214	GP	10/79-4/80	46	0	46	46	37	Officer course No AFQT
EM	6070	GP	8/78-2/79	362	354	349	306	262	
FT	6027	GP	10/78-4/79	77	77	66	66	34	
GM	6115	GP	7/79-12/79	47	47	47	47	21	
HTI (Phil)	6120	GP	10/78-3/79	319	318	319	0	218	Same graduates in CDP 6339 No EOCG
HTI (Treasure Island)	6119	GP	10/78-4/79	264	260	198	197	171	Same graduates in CDP 6106
RM-Sea	6380	GP	2/79-12/79	496	495	487	0	369	No EOCG; Data on same graduates collected from RM Basics, CDP 6144
RM-Shore	6381	GP	3/79-8/79	400	397	400	0	396	No EOCG: Data on same graduates collected from RM Basics, CDP 6144
SH	6477	GP	10/79-3/80	177	176	177	177	132	
Subtotals		SP	10/78-4/80	1830	1720	1679	941	1188	SP = Self-Paced
		CMI	3/79-4/80	823	806	803	708	290	CMI = Computer Managed Instruction
		Mixed	10/78-4/79	264	260	261	190	171	Mixed = Self-Paced and Group-Paced
		GP	8/78-4/80	2361	2281	2217	1010	1758	GP = Group-Paced
		Total	8/78-4/80	5278	5067	4960	2849	3407	

[1]Course type based on designation during 1978-1980.

[2]Converted to CMI in FY-81.

[3]Initially based on school records and/or SME reports.

Data Collection

The names, graduation dates, and SSNs of school graduates were obtained from NAVEDTRACOM TAS files. Visits were made to the schools between August and November 1980 to obtain data on course graduates. At the schools, data were manually recorded from class records and entered on worksheet forms for subsequent entry into computer files. Data recording was accomplished either by TAEG project staff or school SMEs functioning under general TAEG supervision. Information recorded, consisted principally of end of course grades and time to complete training. Where available, other measures of training effectiveness/-efficiency were also recorded. These other measures included numbers of academic remediations and setbacks, and numbers of additional hours of instruction required Cost data were obtained from CNET.

Training adequacy ratings and questionnaire task statements were obtained from the CNET Training Appraisal System (TAS) Data Bank. The data included the fleet supervisors' TAS ratings for each graduate (i.e., 1, 2, 3, 4 or 5) on each skill/knowledge item of the course feedback questionnaire, and the mean TAS rating (computed over all items of the questionnaire) for each graduate of the course. The twelve ASVAB subtest scores of each graduate were also obtained from the CNET TAS data base (or the student master file when necessary). During school visits, assigned SMEs classified the items on the TAS questionnaires into the generic task categories. Their classifications were also entered into the data base.

Data Analysis

Analyses of training effectiveness and training efficiency data were conducted on four different groupings of the courses in the study. These groupings are shown in table 3.

First Analysis. The first analysis was based on data of 19 single-phase enlisted "A" schools. The purpose of this analysis was to investigate possible differences between conventional and individualized instruction across courses as well as to assess interrelationships with the two other predictor variables (i.e., student ability level and type training task).

Second Analysis. The second analysis was based on data of two dual-phase enlisted "A" schools. In these schools, enlisted personnel received a different method of instruction in each phase of the "A" school. A major purpose of the second analysis was to investigate the possibility of transfer effects between methods of instruction within the same group of graduates.

Third Analysis. The third analysis was based on data of two single-phase officer courses. Both officer courses nominally presented similar subject matter but each under a different method of instruction. Thus, the purpose of the third analysis was to investigate possible differences in methods of instruction between the two officer courses when training content is held "constant."

Fourth Analysis. The fourth analysis was based on data obtained from all the courses included in the previous three analyses plus three basic, pre"A" school courses. The purpose of the fourth analysis was to provide an in-depth investigation of training efficiency and/or effectiveness within each course studied.

The findings from the first analysis are presented in the subsequent paper in this volume by Freda, Hall, and Ford. The findings of the other analyses will be contained in reports currently being prepared by the TAEG.

REFERENCES

Ellis, J. A., Wulfeck, W. W. II, and Fredericks, P. S. The Instructional Quality Inventory: II. User's Manual (Special Report 79-24). San Diego, CA. Navy Personnel Research and Development Center.

Orlansky, J. and String, J. Cost-effectiveness of Computer-based Instruction in Military Training (IDA Paper P-1375). Arlington, VA: Institute for Defense Analyses, April 1979.

TABLE 3. DESCRIPTION OF COURSES INCLUDED IN THE FOUR ANALYSES

FIRST ANALYSIS				SECOND ANALYSIS				THIRD ANALYSIS				FOURTH ANALYSIS			
Name	CDP	N	Type	Name	CDP	N	Type	Name	CDP	N	Type	Name	CDP	N	Type
AD	6501	464	CMI	HT2	6339	319	SP	DCA	3214	46	GP	All courses in previous analyses plus			
EN	6487	359	CMI	HT1	6120	319	GP								
AKK	6522	254	SP	HT1	6119	264	GP	DCA	3216	60	SP	RM Basics	6144	394	CMI
DK	6061	110	SP	HT2	6106	264	SP+GP					PE Basics (EN)	6261	359	CMI
IM	6046	18	SP									PE Basics (MM)	6262	552	CMI
MM6	6493	374	SP												
MM12	6492	204	SP												
PN	6102	84	SP												
TD	6521	185	SP												
YN	6057	222	SP												
AW	6515	90	GP												
ASE	6530	36	GP												
AW	6537	47	GP												
EN	6070	362	GP												
FT	6027	77	GP												
GN	6115	47	GP												
RMSea	6380	486	GP												
RNSh	6381	400	GP												
SH	6477	177	GP												

CMI = Computer Managed Instruction
SP = Self-paced
GP = Group-paced
SP+CMI = Individualized Instruciton
GP = Conventional Instruction

RELATING SELECTION PREDICTIONS TO ATTRITION IN THE BRITISH INFANTRY

C L Heaton, Army Personnel Research Establishment,
Farnborough, Hants, UK

The current selection system for non-officer ranks is described with details of previous validation work. Typically low non-significant correlation coefficients were obtained with Infantry groups. Current work in this area is severely hampered by the lack of suitable criteria but strategies for overcoming this are under discussion. Length of service was correlated with selection results for an Infantry group and no correlations could be demonstrated. However, data used show that exit rates are dependent on individual training depots. The effects of civilian unemployment in depressing wastage directly and indirectly (by raising the ability of serving soldiers as a population) are discussed.

RELATING SELECTION PREDICTIONS TO ATTRITION IN THE BRITISH INFANTRY

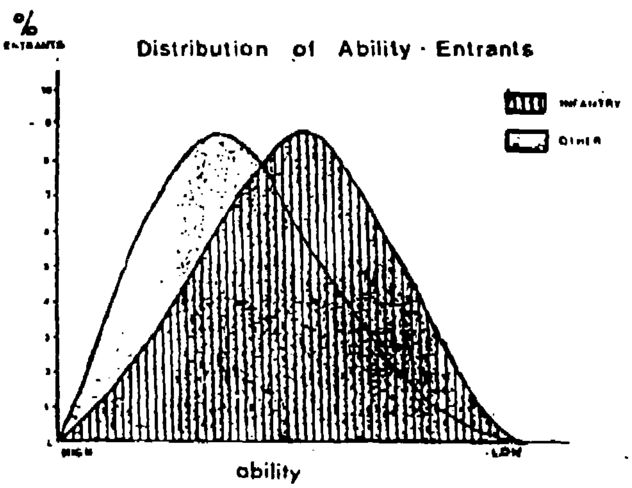
INTRODUCTION

1. The initial aim of this research was to evaluate the validity of the tests currently used to select Infantry soldiers in the British Army. However, due to problems associated with the availability of suitable criteria for these purposes, the data collected, was used for a second purpose, that of examining the wastage rates from two comparable Infantry training depots. Unemployment, a key factor in this instance, is also relevant in any discussion of the recruitment pool and consequently test effectiveness. These three topics, validity, attrition and unemployment are therefore the main areas discussed in this paper.

BACKGROUND

2. The British Army was established as an all-volunteer force from 1960. In recent years, recruitment targets have been set around the 20,000 mark (men, women and junior entry) though, due to the current economic climate targets are currently set much lower to offset a decline in the attrition rate. In 1971 a centralized selection system was set up and recruitment and selection is now seen as a two stage process. Recruitment, career guidance and initial selection are carried out at the local Army Careers Information Offices (there are approximately 170 of these distributed in major towns and cities in the UK). Subsequent allocation to one of the 180 employments for Adult males is made at the centralized selection centre on the basis of results from psychometric tests (pencil and paper achievement and reasoning tests), interview, medical and fitness tests. Final allocation is dependent on there being vacancies to match the applicant's level of ability and expressed choice.

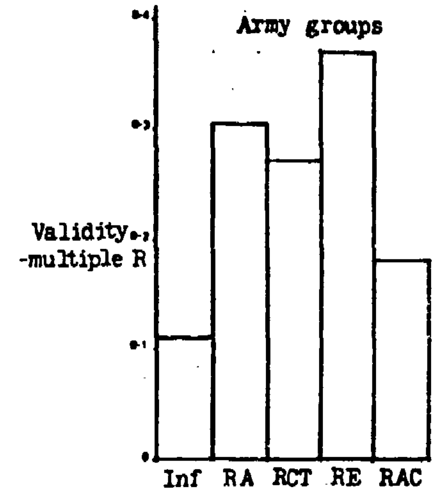
3. Infantry soldier is just one of the employments which may be chosen and subsequently allocated. Approximately one third of a year's intake will be allocated as Infantry soldiers. The selection requirements in terms of test scores required are comparatively low, and the distribution of ability is somewhat skewed, as is shown below.



VALIDATION

4. In 1976 a study was undertaken by Killcross et al to determine the validity of the five battery tests in use at that time. Training criteria were used in this instance. The coefficients produced for different Army groups (Infantry, Royal Artillery (RA), Royal Corps of Transport (RCT), Royal Engineers (RE) and Royal Armoured Corps (RAC)) are shown below.

Validity of Battery tests with different Army groups



As can be seen the validity coefficient for the Infantry is lower than that for the other Arms and Corps investigated. This difference may have been even greater if job performance criteria were used, as one would expect greater degradation of training from lower ability groups such as the Infantry. Job content in the Infantry is also less closely related to attainment type tests than training criteria and possibly job content of more technical trades, another reason shown by Ghiselli (1966) to cause low validity coefficients.

5. Since this work was undertaken, the battery of 5 tests has been revised. All the tests have been changed to consistent multiple-choice format and the item content has been updated somewhat. The complete battery came into use in January 1981. At this point the problems of validating the revised tests were first noted. Having 180 employments, each with different entry requirements, it is both time-consuming and ineffective to manually match selection and training records for each employment. It is essential that information is available for individual training groups (different Arms and Corps undergo training of different types and lengths) as the validity estimate produced for a composite of mixed training groups may be much lower than for the individual training groups themselves (Killcross et al 1976). The use of the Army's computerised manning and records system as a data base for validation is an alternative approach.

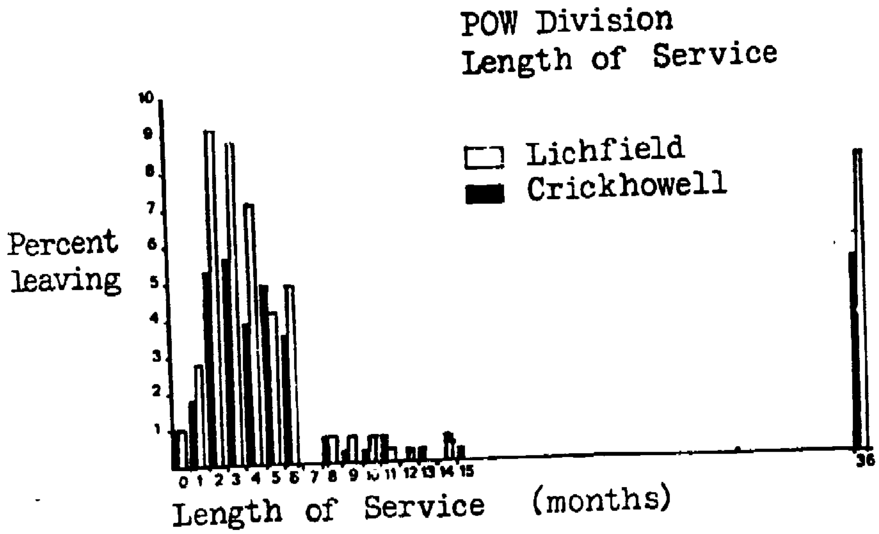
6. A complete listing of soldiers allocated to the Infantry for November 1978 to October 1979 was obtained. As a sample, all those allocated to the various regiments of the Prince of Wales Division were selected. From the computerised system, limited information was then available. This included (in addition to personal details) details of initial selection (an overall grade only, no test-score details), promotion, specialist courses taken and exit dates. As a maximum, a soldier from our sample would have served three and a half years. Few had achieved promotion or specialist qualifications. The only criteria available was length of service. If a soldier has a three year engagement and wishes to leave at that three year point he must give notice at the 18 month point. From our data, we were therefore aware of the soldier's plans to leave, even if he had not yet left.

7. It was not possible, in this instance to use performance on the basic training course as a criterion, as detailed records are not available, and the pass/fail criterion recorded is inadequate as all those soldiers remaining in the Infantry do in fact pass. When the results of individual test scores were correlated with length of service, negligible coefficients were produced ranging from 0.01 to 0.05. This is as would be expected, due to lack of relevance, of the predictor to the criterion (Nagle 1953).

8. As has already been noted, the current economic climate has radically affected the size of the recruiting pool. In the past, when recruiting was more difficult, the selection ratio was very high and for the Infantry was almost equal to unity(excluding those unsuitable on medical or security grounds). Under these circumstances no test will operate with great efficiency and in fact there may be little point in using the tests at all (Taylor and Russell 1939). Now that the selection ratio is more favourable, the validity of the tests becomes more important. Recommendations for changes to the records system, which would allow suitable selection and training data to be stored, are currently under discussion.

ATTRITION

9. The data analysed, did however, show variations of interest. The Prince of Wales Division consists of a number of regiments, each recruited in a different area of the country. Some of the regiments receive their basic training at the Prince of Wales training depot at Crickhowell, South Wales, others at the second Prince of Wales depot at Lichfield in the Midlands. Differences in attrition rates between Divisions of the Infantry has already been demonstrated in an earlier phase of this project. Wastage rates varied from 20 - 50% for different divisions (1979-80 figures). Now, further analysis has revealed differences between depots within a division.

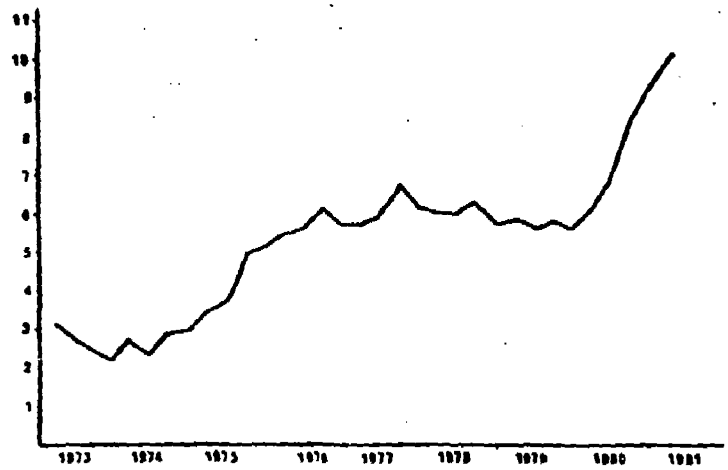


10. The pattern of wastage is, in fact, typical of the British Army as a whole. During the first 6 months of service the recruit has the right to leave the Army either free or on payment of a nominal sum. This period covers the time spent in basic training. After the 6 month point the sum payable is much increased, and few people leave of their own choice until the 3 year point, when their engagement may run out. The majority of the sample studied were in fact on 3 year engagements and most of those who remain after 6 months decided at some time during their first year to extend their engagement.

11. Great differences between the two training depots are also evident. Overall, attrition was far greater for soldiers trained at Lichfield (50% compared with 34% of Crickhowell trained soldiers). There are two areas of importance, firstly, the period covered by initial (basic) training at the depot itself (first six months) and secondly the numbers leaving (or having given notice to leave) at the three year point. In considering training wastage, it is not yet possible to separate the overall effects of external and internal factors. Work carried out in the Netherlands (Tromp, 1981) suggests that some proportion of wastage is dependent on the attitudes and instructional regimes of the training depot staff. These are internal factors. Work carried out in the UK (Dennison 1981) further relates wastage to civilian unemployment, an external factor. However, when examining subsequent attrition, it is likely that, now the training depot is no longer influential, civilian unemployment is a major factor.

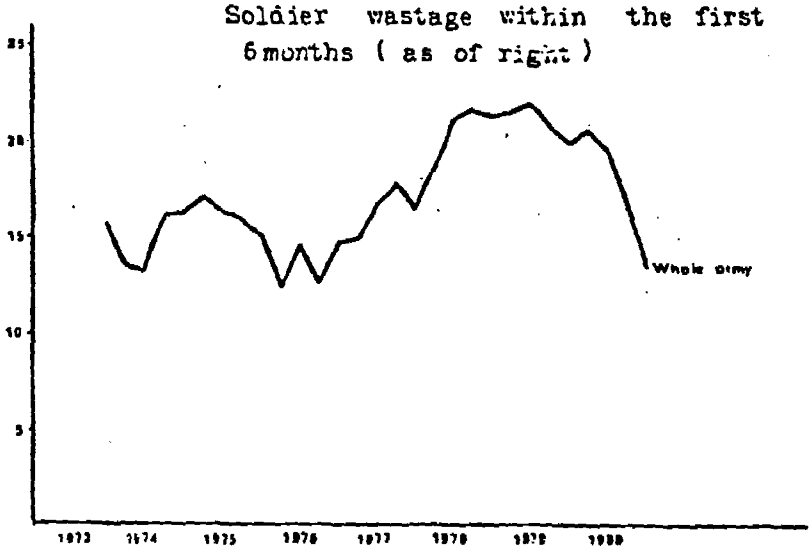
12. The link between civilian unemployment and attrition is obvious - as unemployment rises, fewer people leave the Army as of choice. During 1980/81 unemployment has risen steeply in Britain and currently stands at about 12% of the working population.

Unemployment levels in UK
(% of work force)

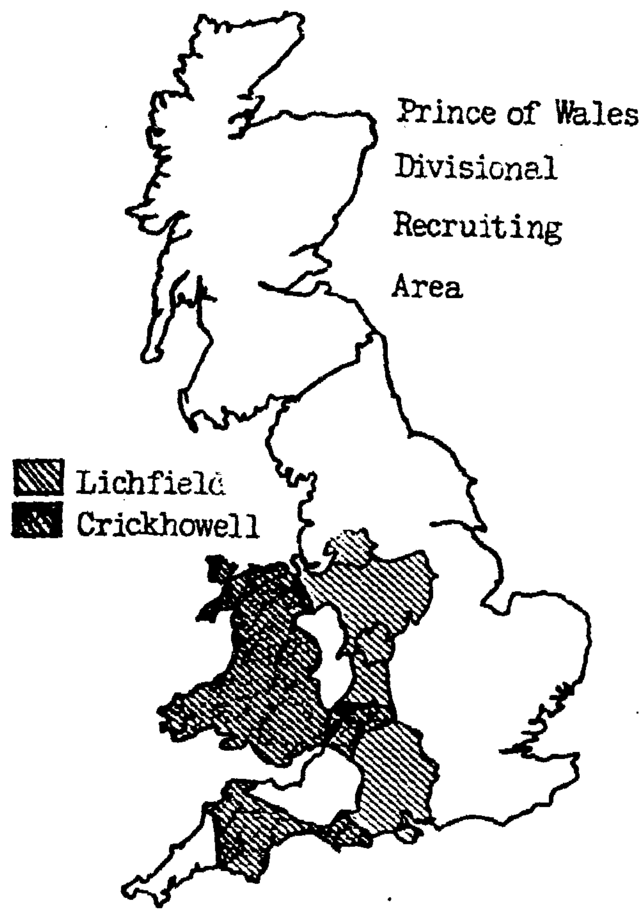


As can be seen, soldier wastage during the initial training period has dropped dramatically over the same period.

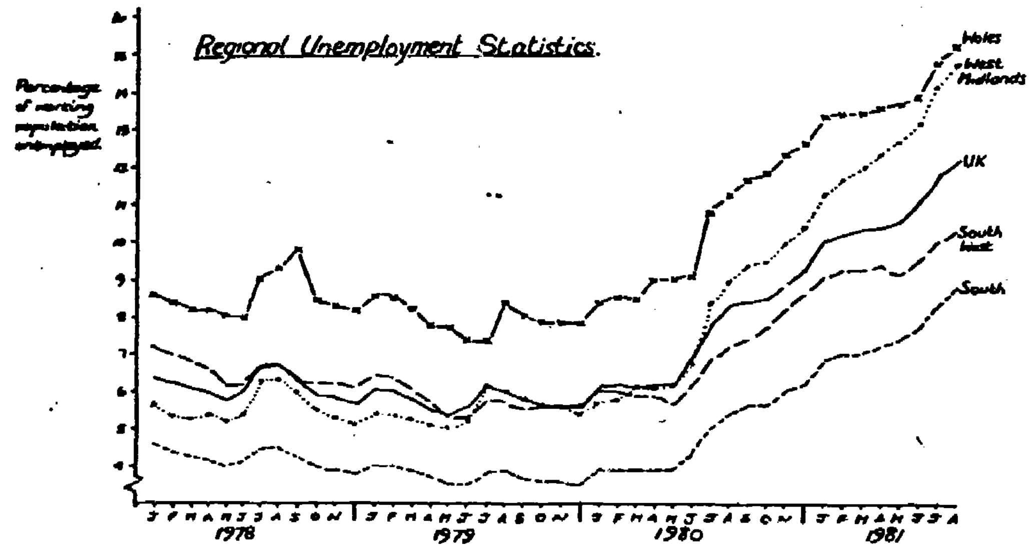
Soldier wastage within the first
6 months (as of right)



13. Wastage is affected both directly and indirectly by unemployment rates. The direct effect is illustrated by the Crickhowell/Lichfield differences. Soldiers originally trained at Crickhowell come from Wales, Devon and Dorset, Avon and Gloucestershire. Lichfield trained soldiers originate from the Midlands and the South of England. These recruiting areas are shown below.



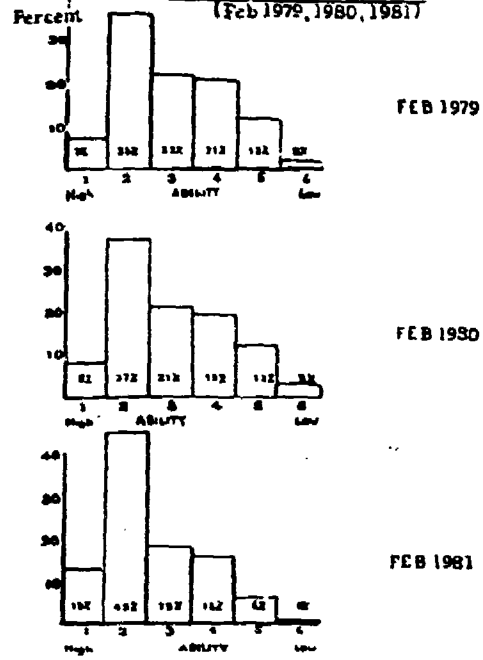
14. These regions have differing levels of unemployment as shown below.



Overall, the Lichfield depot trains men from areas which have experienced lower unemployment over the past 3 years than those areas recruited to Crickhowell. We would expect attrition, especially at the three year point to be closely associated with regional unemployment. This effect would be even stronger if individual regiments, drawn from one area only, were considered.

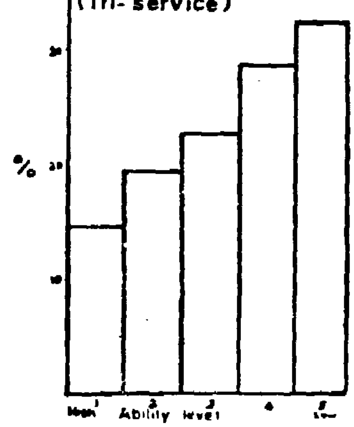
15. The effect of unemployment will be found throughout the initial training period in addition to the internal effects of the depot iself. Unemployment also affects the system indirectly. Increasing unemployment is one factor which improves the quality and quantity of applicants coming forward, as shown by Bellany (1978). This improves the selection ratio and therefore effectiveness of any testing procedure (Taylor and Russell 1939) and increases the proportion of brighter candidates in the recruitment pool. Such a change has been observed over time in the British Army.

ABILITY OF ADULT MALE ENTRANTS
TO THE BRITISH ARMY
(Feb 1979, 1980, 1981)



16. Work carried out internationally by the Technical Co-operation Panel has shown that wastage is greater in lower ability groups. For the British services overall, this can be seen to apply.

Wastage by Ability Level
(Tri-service)



It follows that a higher ability intake will, over time produce fewer losses. This effect, will of course, operate with a time lapse and may now be making its first impressions on the wastage figures.

CONCLUSIONS

17. Though previous research has been problematic and has produced only low validity coefficients, it is important to evaluate and monitor on a regular basis, the validities of selection tests, particularly, as is currently the case, in times of favourable selection ratios. Length of service is an inadequate criteria for this purpose. Unemployment has affected both recruitment to, and wastage from the Army, and local unemployment effects should be taken into account in a model which attempts to explain retention issues, in addition to factors in operation within the individual training depots. It is important to draw methodological conclusions in this instance. A broader approach to validation problems should be adopted as potentially useful data may otherwise remain untapped.

REFERENCES

1. Bellany I (1978). A statistical analysis of factors affecting voluntary enlistment into the UK Armed Services 1960 - 1976. University of Lancaster.

2. Dennison (1981). Relating recruiting and selection predictions to wastage in the British Army. Paper presented to the 17th International Symposium on Applied Military Psychology. Lisbon 1981.

3. Fabyanic T A (1976). Manpower trends in the British All Volunteer Force. Armed Forces and Society Vol 2 No 4 August 1976.

4. Ghiselli (1966). The validity of occupational aptitude tests. New York. John Wiley Inc.

5. Killcross M C et al (1976). The validity of standard and experimental selection measures used in the Pilot Centralised Selection Scheme for Adult Recruits. APRE Report No 35/75.

6. Nagle B F (1953). Criterion development. Personnel Psych 6 271-289.

7. Taylor H C and Russell J T (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and Tables. Journal Applied Psych 23 565-578.

8. Tromp Th H J (1981). Remarks and observations on the problems of adaptation concerning life in the Armed Forces. Paper presented to the World Psychiatric Association (Military Psychiatry Section), symposium on Adaptation to Military Service. Toulon 1981.

Hicks, Jack M., US Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia. (Tues. P.M.)

Counter-Attrition Programs in the US Military

AD P001329

Of special interest were "counter-attrition" programs in the US Military, whether in the conceptual, research, development, or operational stages. This review included the US Army, Navy, Air Force, Marine Corps, and Coast Guard. Attention was addressed to any strategy or mechanism intended to reduce undesirable attrition in a military environment. Both up-front and downstream programs were scrutinized. Programs addressing such major issues as realistic expectations, transition from training to unit, and unit disillusionment were considered. Of interest were pre-military preparation, coping skills, behavioral modification, continuing education, retraining, motivational adjustment programs, counseling, use of post facilities, unit cohesion strategies, and the like.

The primary data gathering tool was the personal and telephone interview. Considering the exploratory nature of this investigation, the interviews were of an unstructured nature. Maximum latitude was permitted to allow for expression of strongly held opinions, and relevant sources of expertise.

Paper presentation at the Annual Meeting of the Military Testing Association, Arlington, VA, October, 1981.

COUNTER-ATTRITION PROGRAMS IN THE MILITARY

Jack M. Hicks
U.S. Army Research Institute for the Behavioral and Social Sciences

High first-term enlisted attrition rates continue to be a major concern in the U.S. Military. The purpose of this paper is to create increased awareness of three categories of military programs which either presently demonstrate, or show promise of impacting significantly upon the enlisted attrition. Programs will be described which represent (a) pretraining, (b) realistic expectations, and (c) retraining.

The military pretraining programs to be represented are not primarily designed to counter attrition. They are programs conducted by the Army National Guard which combine Guard recruitment with job preparation and placement. These programs, however, show sufficient promise in lowering attrition to warrant representation in this discussion.

The first of such programs began in Oakland, California approximately 4 years ago. It was, and still is, designed for unemployed, and economically disadvantaged youth, 17-22 years of age. The basic attraction to the Program is the enhanced opportunity for a job. In order to get the help needed to acquire that job, one must graduate from a 10 week military preparation program and sign for a six year Guard obligation. The greatest emphasis in the Program of Instruction (POI) is given to military skills and basic literacy skills, but also career assessment and pre-employment training. The military skills component is primarily physical fitness training, drills and ceremonies, and military field trips. Basic skills training consists mostly of reading, writing, and mathematics, but with some "survival training." Survival training places a heavy emphasis upon the importance and value of community awareness and service, "banking," "nutrition," "cultural studies," and other fundamental education. Career assessment takes up one (1) week of the program, and consists of testing, evaluation, and vocational counseling. Pre-employment training occupies about 1½ weeks and addresses a plethora of concerns such as how to complete an application form, how to prepare a resume, how to conduct oneself in an interview and on the job, and introduction to selection tests used by employers, and the like. The Program focuses upon the development of coping skills for employment and military settings, discipline, civics, teamwork, and the skills to pass the Armed Services Vocational Aptitude Battery. All of this adds up to $1500 to $2000 per pupil. Perhaps as few as 20% of the applicants are accepted. Of this 20%, approximately 50% graduate. Rejections and failures derive primarily from substandard literacy skills, and negative attitudes toward the military. At least 90% of the graduates successfully meet Army screening criteria, and proceed to Basic Training. Of interest for the active Army is that there has been only about a 10% attrition rate among such pretrained recruits throughout Basic and Advanced Individual Training (AIT). Non-pretrained recruits have attrited at approximately double that rate during the full training phase. Also of potential interest is that perhaps as many as 1/3 indicate that they would prefer to remain with the active Army rather than return to their communities and fulfill their ARNG obligation. At present, upon successful completion of Army training, the new Guardsmen return to their community where they will receive help in getting MOS-related jobs, and attend monthly drills, and summer training. To private sector employers, perhaps the most attractive feature of the Program derives from the credibility of military training. Employers appear to be more willing to hire a youth who has demonstrated the discipline assumed to be required to successfully complete Basic and AIT. At present the original Oakland program has been adapted to Los Angeles and Sacramento, California, Seattle, the tri-city area in Washington state, and Milwaukee, Wisconsin.

Realistic Expectations. The military has made several recent audio-visual efforts specifically designed to provide new recruits with a heightened awareness of what the military will be like and how to deal with it. It has been hypothesized that enlisted attrition should in some degree be inversely related to such increased awareness. The first of these efforts was conducted by the Marine Corps (circa 1978). An 80 minute color video tape was developed jointly by a private contractor and the Training Support Center of the USMC Parris Island Recruit Depot. The film is considered a Realistic Job Preview (RJP), attempting to provide recruits an accurate portrayal of what life in the Marine Corps would be like. It includes footage from all phases of recruit training complete with voice overs and interviews. The film was evaluated by Horner, Mobley and Meglino (1979) based upon its viewing by 678 Parris Island male, first-term, enlisted recruits. It was shown after swearing in, during the second full day at the Recruit Depot. Results did not show a signficant association between film viewing and attrition by the end of training, but did so in the hypothesized direction after 6 months and 12 months.

It would be of interest to have the evaluation of the above film replicated by an independent research group. Such an independent evaluation was conducted in regard to a San Diego (Navy) version of the Marine Corps film. The San Diego film involved the same contractor, and was produced by the Center for Naval Technical Training in 1979. It is approximately 1 hour in length and shows actual Navy enlisted recruits, DIs, chow lines, obstacle courses, and the like. No professional actors were used. The evaluation of this film showed no significant differences in attrition rates between those (n=1049) who viewed the film and those (n=1002) who did not (Lockman, 1980). Similar "no impact" findings have been reported by a recent initial evaluation of a Great Lakes version of this film.

An additional Navy effort is a 20 minute recruit "coping skills" film (Sarason, circa 1979; personal communication). The film content focuses upon "(a) what to expect, (b) how you may feel, and (c) what to do." Again, initial evaluation has shown no impact upon attrition. A 6 month follow-up of this film is forthcoming.

More recently, the Army has developed realistic expectations and coping skills films. The realistic expectations film is a 28 minute color video-tape providing information about the content of Basic Combat Training (BCT). It consists of four components: Introduction, Weapons Training, Individual Tactical Training, and Necessary Testing Activities (I WIN). This videotape, like the others, shows actual situations, though it concentrates on training. Though it has a professional narrator, all participants are actual on-site personnel at Ft. Jackson, South Carolina. At present, the film is operational at all Army installations where purely BCT is offered. It is either shown within the context of the reception center or slightly later in the BCT unit. It has only been very recently that this film was made operational, and no evaluations are available.

Retraining. Of particular interest in this paper are enlisted retraining programs. These are basically interventions which perhaps represent the closest device the U.S. Military presently employs toward the establishment of counter-attrition programs, per se.

Of primary interest are programs designed for marginal performers and/or problem personnel who show potential for development into productive soldiers or sailors. The most comprehensive and established of such programs at the present time is the U.S. Army Retraining Brigade (USARB) at Camp Funston, Ft. Riley, Kansas. The original USARB Program, known as the Army Correctional Training Facility, became operational in 1968. It represented a landmark for the Army correctional system due to its emphasis "upon systematic restoration and utilization of potentially wasted manpower." (USARB Annual Report, 1980). During its first 4 years, the Correctional Training Facility received over 23,000 prisoners from Army stockades worldwide and returned over 19,000 for subsequent service.

Physical training is the feature most emphasized in the present USARB POI. It is deliberately designed to exert sustained physical and mental stress within a spartan military environment. Mental stress is generated from continued observation, daily evaluations on a variety of dimensions, peer pressure, and high performance standards.

USARB graduates are given a special Enlisted Evaluation Report (EER) after 60 days in their new duty assignments. Of 457 FY80 graduates returned to duty, 71% were rated as promotable immediately or ahead of their peers. Only 10% were rated nonpromotable. This seems a remarkable turnaround considering that only about 4 months previously, these individuals were awaiting court-martial in an Installation Detention Facility (or stockade).

A significant outgrowth of USARB, is a co-located program at Ft. Riley called the Individual Effectiveness Course (IEC). This program became operational in 1977 and is significant in that it has served as a model program for light offenders who are regarded as having potential for productive duty. The IEC is distinguished from USARB in that it is for Ft. Riley enlisted personnel only, is a non-confinement facility, and is designed to counteract maladaptive propensities before significant punishable offenses actually occur. The POI is 6 weeks long and otherwise is very similar to USARB. There may be as many as 50 trainees in each class, who may or may not graduate. Recommendations as to retention or elimination are made to the unit commander in each individual case.

In a study of 197 IEC trainees, 54% graduated and returned to duty, 40% were recommended for various administrative separations, and 6% were lost through AWOLs or medical profiles. At initial referral, unit commanders indicated probable actions to be taken within 30 days if the soldiers had not been admitted to the IEC. Predicted actions were 33% Expeditious Discharges, 32% Chapter 13s (unsuitability), 27% miscellaneous transfers and administrative separations, and 8% Article 15s. That is, some punitive action was expected for all 197 trainees. Fifty-four percent of these 197 graduated from the program, and returned to duty. As with USARB, EERs were conducted for these graduates 60 days after returning to duty. In terms of advancement potential, nearly half were rated as promotable immediately or ahead of their peers. This seems to stand in marked contrast to the above dim prognosis prior to IEC training.

A counterpart Navy program became operational in 1979 after having carefully studied USARB and the IEC. This program, called the Behavioral Skills Training Unit (BEST) is located at the Naval Amphibious Base in Little Creek, Virginia. It is shorter in duration than the IEC - 4 weeks instead of 6--but again heavily emphasizes physical training such as daily distance running, an obstacle course twice weekly, and group sports. Additional emphasis is placed upon individual and group counseling, goal setting, and academic advancement. The mission of BEST is to provide behavioral skill training to low and marginal performing first term enlisted personnel that will enable them to successfully complete their obligated service. Criteria for admission include no disciplinary action pending, at least 2 years of active obligated service remaining, and the potential to complete enlistment, but unlikely to do so given the present demeanor and record of achievement.

As of June 1981, BEST had enrolled 47 classes, averaging approximately 24 each, for a total of 1145 trainees. Of these, 86% graduated, with the remainder either failing the program or being returned to their Command. All participants, including nongraduates, are evaluated at 6 and 12 month intervals subsequent to BEST training. Of 700 6 month evaluations, 69% were rated as performing average or above. Six months prior to BEST training, 78% were awarded non-judicial punishments and/or courts martial, whereas 6 months after BEST, only 33% received similar punishments. Twenty-two percent have received recognition for outstanding performance, and an additional 22% have been promoted at least once. The 12 month evaluations reveal a similar pattern, with even greater percentages (79%) having received recognition for outstanding performance or at least one promotion. Again, as with USARB and the IEC, a dramatic turnaround seems indicated.

An additional Army program clearly shows promise as a potentially productive counter-attrition program. This program, called the Intensified Training Unit (ITU) is co-located with the Installation Detention Facility (stockade) at Ft. Carson, Colorado. The ITU seems unique at the present time in that it may be the only Army Correctional Custody Facility (CCF) dedicated to a retraining rather than a work detail orientation. Again, USARB and the IEC served as models for the development of the ITU POI. The POI is 30 days, and is consistent in content and objectives to programs already described. The ITU received its retraining mandate in 1978 as a part of an effort by the Ft. Carson Law Enforcement Command to upgrade its correctional program to a more remedial as opposed to custodial format. Like other CCFs, the ITU is a non-confinement unit, providing for close supervision over soldiers who commit minor infractions. A key limitation, at present with the ITU, is that a punishable offense is required for admission. A review of statistical data maintained for a three month period in 1978 showed that of 67 correctees who completed the ITU and returned to their units, 42 have since received performance ratings of good to outstanding. More substantial data are needed for a complete evaluation of the effectiveness of the ITU, but preliminary indications suggest positive outcomes similar to previously described programs.

Other similar programs could be mentioned that now exist, or are getting under way, throughout the Armed Forces. The above programs are among the oldest, best established, best evaluated, to date. There are indications that such programs are proliferating at the local level. The ITU, for example, has sent packets of information to at least a half dozen interested installations thus far. Both the Defense Audit Service and the Army Audit Agency have made their own analyses of several of the above programs--both concluding cost-effectiveness and recommending expansion.

TREND ANALYSIS AND PROJECTIONS OF THE POPULATION AND RACE/SEX COMPOSITION OF THE FEDERAL CRIMINAL INVESTGATING SERIES

Hannah Rothstein Hirsh
Personnel Research Psychologist
Personnel Research and Development Center
Office of Personnel Management
Washington, D.C. 20415

Data on the Federal Criminal Investigating Series were taken from the Central Personnel Data File (the main repository for employment history and demographic information of Federal civilian employees), for 1977-1980. Four year trends in the size and race/sex composition of the full-time series population were developed. It was found that the total population size was rather stable during the past four years, but that various race/sex groups experienced considerable growth or decline.

The trend data were used as the basis of projections for the 1981-1985 period. Among the assumptions upon which the projections were based was that the 1981-1985 period would be similar to the preceeding four years; probable divergences were pointed out, and their effects speculated upon.

The usefulness of data such as those examined in this paper for purposes of forcasting and human resources planning was considered, and determined to be positive, if incomplete.

Data Source

Except where otherwise noted, figures in this report are based on the entire population of criminal investigators working in the United States during 1976-77 to 1979-80. Records of all full-time permanent employees of the criminal investigating series contained in the Central Personnel Data File (the main repository for employment history and demographic information of civilian employees of the Federal Government) were included in the data analyses. Thus, obtained figures are more properly regarded as population parameters than sample statistics. Among other things, this renders the use of tests of statistical significance largely inappropriate.

It should be noted that any yearly figure is as of June 30 of that year; a one year period is from July 1 of a year to the next June 30.

This paper focuses on total population and on race/sex composition, but it is part of a larger study which also examines other demographic characteristics ,and personnel dynamics such as accessions, losses and inter-occupational movement.

Employment in the 1977-1980 Period

Population Changes

Total Population. Data from the Central Personnel Data File (CPDF) indicate that the number of full-time permanent employees in the criminal investigating series changed very little in the three years from June 30, 1977 to June 30, 1980. As Table 1 shows, the number of full-time permanent employees on board on June 30, 1980 was 19,064. This is less than one percent larger than the June 30, 1977 figure of 18,920. Changes across one year intervals were never greater than 1 1/3 percent (which was the amount of increase between June 30, 1979 and June 30, 1980) . The average yearly change in the total-full time permanent population of Federal criminal investigators was .28%.

Table 1

Full-Time Permanent Employees in the Criminal Investigating Series, 1969-1972 and 1977-1980

Series Totals

June 1969	June 1970	June 1971	June 1972	June 1977	June 1978	June 1979	June 1980
14,015	15,196	16,834	18,852	18,920	18,734	18,806	19,064

Percent Change

1969-1970	1970-1971	1971-1972	1969-1972	1972-1977	1977-1978	1978-1979	1979-1980	1977-1980	1969-1980
8.4	10.8	11.9	35.0	.36	-.98	.38	1.37	.76	36.0

Table 1 also compares the number of full-time permanent employees in this series during the three most recent years to the series population during the period June 1969-June 1972 (U.S. Civil Service Commission, 1973). The two sets of figures are not exactly equivalent since the data from the 1969-1972 period is an estimate based on a sample of all General Schedule employees in the criminal investigator series. The small number of non-permanent full-time employees, however, increases the comparability of the two sets of figures. It can be seen from the information in the table that the years 1969-1972 were years of growth for the criminal investigator series, with an average annual increase of 10.4%. The difference between 1972 and 1977, however, is .36%, or 1/3 of one per cent; and the change in the criminal investigator population between June 1969 and June 1980 of 36% is only one per cent greater than the 35% change between June 1969 and June 1972. These data provide support for the claim that the total population in series 1811 has been relatively stable for the past few years, although it is possible that fluctuations in different directions between 1972 and 1977 balanced out to present a picture of seeming lack of change. (Directly comparable data, i.e. series totals as of June 30 1972-77, are not available. The series population totals which are available, as of October 31 of the relevant years, cannot be compared to the June figures because of seasonal differences).

Sub-groups. Although the population as a whole displays considerable stability, there is a good deal of variability in the numbers and proportions of different race/sex sub-groups. Table 2 presents these data for the years 1977-1980. It is obvious that the occupants of the criminal investigator position are overwhelmingly non-minority (91.7% in 1980) and male (96.1% in 1980). Nevertheless, an increasing number of women and minorities have entered this occupation since 1977, while the proportion of non-minority males has decreased by about 4%.

As Table 2 shows, in 1977 the population of criminal investigators in the Federal service was 98.4% male, and 94.1% non-minority. By 1980 these figures had decreased to 96.1% and 91.7% respectively. There has been a slight decline in the proportion of males each year, as well as in the proportion of non-minority employees, due entirely to a decline in the proportion of non-minority males. From 1977 to 1980, there has been an overall increase in the number of females of 150%; the average yearly increase was 35.92%. The total proportion of women in this occupation, however, remains low-- 3.9% as of June 1980. The increase in non-minority females has been 152.55% overall, with a yearly average increase of 36.41%; their present total proportion is 3.4% of the total series population. The increase for minority females has averaged 32.77% annually, for a total increase from June 1977 to June 1980 of 135.71%. The current proportion of the series population represented by minority women remains quite small-- about .5 percent. Total minority representation has increased by 42.95% in the period 1977-1980, averaging 12.73% a year. The current proportion of minorities is 8.3%. The increase for minority males is also consistent with this general trend. The overall increase was 39.31%; the average yearly increase 11.77%. The proportion of minority males in the total series has increased from 5.9% to 8.3% between June 1977-June 1980.

Table 2

Proportions of Full-time Permanent Criminal Investigators Belonging to Various Race/Sex Groups, 1977-1980

Series Population

	Number				Per cent of Population*			
	1977	1978	1979	1980	1977	1978	1979	19
Male	18,623	18,319	18,281	18,321	98.4	97.8	97.2	96
Female	297	415	525	743	1.6	2.2	2.8	3
Non-Minority	17,812	17,550	17,422	17,480	94.1	93.7	92.6	91
Non-Minority Male	17,557	17,192	16,973	16,836	92.8	91.9	90.2	88
Non-Minority Female	255	358	449	644	1.3	1.8	2.4	3
Minority	1,108	1,184	1,384	1,584	5.9	6.3	7.4	8
Minority Male	1,066	1,127	1,308	1,485	5.6	6.0	7.0	7
Minority Female	42	57	76	99	.2	.3	.4	

* Percentages may not add up to exactly 100% due to rounding

Per cent Change

	1977-1978	1978-1979	1979-1980	1977-1980
Male	-1.63	-.21	.22	-1.62
Female	39.73	26.51	41.52	150.00
Non-Minority	-1.47	-.73	.33	-1.86
Non-Minority Male	-2.08	-1.27	-.81	-4.11
Non-Minority Female	40.39	25.42	43.43	152.55
Minority	6.86	16.89	14.45	42.96
Minority Male	5.72	16.06	13.53	39.31
Minority Female	35.71	32.33	30.26	135.71

Employment Projections to 1985

About Projections

Projections are used to provide information about the future growth or decline of a population. A projection does not necessarily represent the "best guess" about, or most probable estimate of, a population at a given point in the future; rather, it is a picture of the population based on a set of assumptions and hypothesized relationships which are more or less likely. Thus, some projections may be quite unlikely to occur, but are made to illustrate a "what-if" situation of theoretical importance; others are based on what the author of the projection believes to be the most probable future occurrences. This latter class of (most probable) projections are properly termed forecasts.

Because the accuracy of assumptions is not always certain, and because the assumptions underlying a projection have such a great effect on the projection, it is considered desirable to develop more than one set of projections of population change. For example, a series of projections recently released by the Bureau of Labor Statistics (Fullerton, 1980), displayed three labor force growth scenarios for the period 1985 to 1995. High, intermediate and low growth projections were made which reflect different overall participation rates for males and females, blacks and whites, for the four major race/sex sub-groups, and for different age groups. The participation rates assumed for each scenario were, in turn, based upon three (high, middle and low) series of projections of general population change produced by the Bureau of the Census.

For general population projections there are three major components of population change about which assumptions are typically made: fertility (births), mortality (deaths) and migration (population movement into and out of the area). Projections may reflect assumptions about how these factors will change, combine or intereact to affect the future population. For labor force studies, such as the present, assumptions may be made about accessions, separtions and migrations between jobs or occupations, as well as about total labor force participation or size.

Assumptions and Components of Projections in the Current Study

Because of severe limitations in the data resources available to the current project, it was possible to make only very simple projections of future changes in the population of the 1811 series. The projections of series and sub-group populations and of their demographic characteristics are based on the assumption that the trends of the recent past (1977-1980) will continue during the period of projection (1981-1985). In other words, it was expected that the rate of change in the population during 1981-1985 will be similar to that experienced during the past four years. Furthermore, contributions of changes in each of the three major components of population change (namely accessions, separations and occupation changers) are not used as the bases for assumptions about total series population change. Rather, total population change is projected on the way the total population changed in the recent past.

Accessions, losses and occupation changers have been examined individually in another part of this study and projections of their future growth are made. These data are not, however, then used to "correct" or revise the population projections presented in the first half of the report. This is one limitation of the present study.

A major risk involved in using average growth in the period 1977-1980 as the basis for projections for 1981-1985 is that the base period may have been atypical or unusual in some way. For example, the active recruitment of females and minorities during the base period may have lead to a higher proportion of female and minority criminal investigators than would be the case under other conditions. Active minority and female recruitment may not continue through the 1980's. A parallel but opposite risk is that the period of projection may be one that departs from even long-standing trends in the target occupation because of changes in the external environment. Thus, severe budget constraints, which appear to be the trend of the future, may interfere with the past pattern of population change. To the extent that the 1981-1985 period differs in any component of population change from the 1977-1980 period, the projections made in this report may diverge from actual trends.

Development of Projections

Two methods of making simple linear projections of series population and race/sex composition based on the CPDF data reviewed above, were tried. One method, which was based on the average annual proportion of change, resulted in a bloated population figure, and severe disparities in projected population size depending on whether the projection was based on growth of the total series or growth of individual race/sex sub-groups . This method was rejected as unrealistic. The other method, which was based on the change in average annual population frequencies, produced a scenario with slow overall growth and more rapid, but contained growth in the number of minority and female criminal investigators. This second method was viewed as more realistic, and was used as the basis for further projections. The two methods and their outcomes are discussed in more detail below.

The Proportions Method. Figures projected by this method are based on the assumption that changes in population in the series, and among its subgroups included in the report will continue to increase or decrease by the same percentage as in previous years. Using this method results in two widely different figures depending upon whether the series population projected for 1985 is based on the average annual proportion increase in the series as a whole during 1977-1980, or based on the sum of projections made for the amount of growth in each of the four major race/sex sub-groups. Projections for each of the sub-groups are derived from their individual average annual proportionate growth. The projected population size, based on the increase in the series as a whole is 19,314. The projected population size for the total population built up from projected increases in each of the four sub-groups is 24,099. The sum of the parts are rather anomalously larger than the whole. This is obviously extremely problematic, and argues against using this method for making projections.

The reason for the discrepancy is that the population as a whole had a low average proportionate increase, but the various sub-groups had widely different growth rates which this method does not "force" to stay within a limit set by the total. In particular, minorities and females had rather high rates of growth, while non-minority males had a low rate of decline. Since the minority and female sub-groups are so small, even a slight numerical increase results in a rather spectacular proportionate increase. For example if, in 1977 there were only 100 female investigators, an increase of 25 new female investigators represents an increase of 25 per cent. The next year, the expanded base of 125 female investigators would require about 30 women to achieve a 25% increase; the third year's base of 155 would require about 37 women to achieve a 25% increase and so on. Thus, as the base population expands, the same numerical increase each year would yield a declining rate of increase; to sustain the same rate of increase requires ever larger numerical growth. As the numbers of females and minorities begin to swell, it becomes more unlikely that they will actually sustain the rate of growth predicted by this projection. It is also extremely unrealistic to expect the series to reach a size of 24,099 by 1985; this is a 27% expansion which is not consistent with any of the other projections in the report. It is more likely to assume that real growth will be closer to the figure of 19,314 calculated from the single total population figure, and that the potential growth in any individual sub-group will be necessarily limited by total series population growth. Figures generated by this method are shown in Table 3.

This method was rejected as a basis of making projections of series change because the resulting figures were so problematic and unrealistic. The data produced from this method of calculation, do however, serve one function. They indicate that it is extremely unlikely that the present rate of growth of minorities and women in the 1811 series can continue unabated.

Table 3

Discrepancies between Projections of Criminal Investigator Population when Calculated for Total Series Population, or as a Sum of Sub-groups, based on Proportions Method

	Non-Minority Male		Non-Minority Female		Minority Male		Minority Female	Total from Sub-groups		Total Calculated Indep.
Average Annual % Increase 1977-80	-1.37		50.85		13.10		45.24			.26
Projection of Pop. to June 1985	15,715	+	4,993	+	2,749	+	642	= 24,099	≠	19,314

The Frequencies Method

Total population. This series of projections is based on the average numerical change in population or sub-population over the past three years. The average annual frequency by which the series, and each of its race/sex sub-groups increased or decreased during 1977-80 is projected as the future annual increase or decrease.

For example, female criminal investigators have increased at an average of 149 per year for the past three years (although the actual numbers have varied from year to year). To make the projection, the number of females in 1980, which was 743, was used as the base. For each year included in the projection 149 females were added. Thus, it is projected that in 1981 there will be 842 females, in June of 1982 there will be 1041 females, and by June of 1985, there will be 1488 female criminal investigators. Similar projections were made for all four race/sex sub-groups included in this study. As Table 4 shows, if present trends continue, it is projected that the 1985 population of criminal investigators will total 19,304. This is less than 1% greater than the June, 1980 series population of 19,064 . Males will make up 92.3% of this group, of which 81.0 % will be non-minority males. Of the projected 7.7% females, it is estimated that only 1% will be minority females.

This method produced projections which appeared to be of realistic size and internally consistent. The figures it produced were therefore accepted for use in this project. This should not be taken as meaning that this set of projections represents the most probable future events; their probability or likelihood in comparison to projections others than those produced by the proportions method discussed above has not been assessed.

Table 4

Projections of Criminal Investigator Population by Method of Frequencies 1981-1985

	Average annual increase 1977-1980	June 1980	June 1981	June 1982	June 1983	June 1984	June 1985
			(P R O J E C T E D)				
Total	48	19,064	19,112	19,160	19,208	19,256	19,304
Male	-101	18,321	18,220	18,119	18,018	17,917	17,816
Female	149	743	892	1,041	1,190	1,339	1,488
Non-Minority	-111	17,480	17,369	17,258	17,147	17,036	16,925
Non-Minority Male	-240	16,836	16,596	16,356	16,116	15,876	15,636
Non-Minority Female	131	644	775	906	1,037	1,168	1,299
Minority	159	1,584	1,743	1,902	2,061	2,220	2,379
Minority Male	140	1,485	1,625	1,765	1,905	2,045	2,185
Minority Female	19	99	118	137	156	175	194

Conclusions

This paper examined in detail two of several important trends in the demographic characteristics and personnel dynamics of the criminal investigating series during 1977-1980 which were studied by the author. It is unfortunate that space and time limitations precluded an analysis of the other demographic and personnel dynamics trends in this forum, but it is hoped that the few data which were presented capture the flavor of the larger study from which they were taken.

It is believed that the systematic reporting of detailed Federal single-occupation data is not duplicated elsewhere. Data on the demographic characteristics of a series population, such as age, education and grade, as well as on population size and race/sex composition, provide a historical snapshot of the population for the period studied. In the present case we have a picture of what the criminal investigating series looked like on June 30 of each year from 1977 to 1980. Personnel dynamics data on accessions, losses and inter-occupational movement, complete the picture by presenting a panorama of the changes in the population during the period of study. It is important to realize, however, that data such as those presented here are a faithful representation of the actual composition and history of the population to the extent that the data bases on which they are based are themselves accurate. The accuracy of the present data bases, namely the Current Status File and the Transaction History File of the Central Personnel Data File, are dubious for some purposes, such as career tracking (Hirsh, 1981), but appear to be sufficiently accurate and reliable for relatively macro planning purposes such as budgeting or examination planning.

Knowledge of the demographic characteristics of an occupation's population as well as of fundamental characteristics of population change is a necessary component of any workforce planning effort. It is also critical for making informed policy decisions in other personnel related areas. This is true whether one wishes to maintain or change current trends in the population of the occupation. It is therefore intended that,in the future, we will continue to analyze Federal-wide single occupation trends, and attempt to improve the ways in which the resultant data can be used in the service of human resources planning and decision making.

References

1. Hirsh, Hannah R. Trend Analysis and Projections of the Demographic Characteristics and Personnel Dynamics of a Single Federal Occupation. Washington, D.C.: Office of Personnel Management, 1981.

2. U.S. Civil Service Commission. Current Federal Workforce Data. Washington, D.C., 1973.

The Strategic Weapon System Training Program
Part II-Executive Steering Group's Role -- SWS Personnel and Training Evaluation Program

Edward N. Hobson, Ph.D.
DATA-DESIGN LABORATORIES

ABSTRACT

This paper describes how an Executive Steering Group (ESG) built a "management team" consisting of members from three organizations primarily responsible for the evaluation and feedback component of the Strategic Submarine Force SWS Training Program described in Part I of this two-part presentation. PTEP (Personnel and Training Evaluation Program) is the component within the SWS Training Program which has a broad evaluation and feedback charter.

The ESG addresses policies, issues, and activities as they relate to the whole PTEP evaluation responsibility within the training system. The variety, scope and complexity of PTEP evaluation repsonsibilities that must be carried out in a dynamic environment require close coordination, frequent communication and periodic assessments/status reports from those assigned major tasking.

This paper discusses the dynamic environment and the factors at work when ESG meetings were instituted, describes the composition of the group, tells how actions are assigned and tracked, and the contribution of the ESG to developing and maintaining a functioning "management team" in support of the SWS Training Program.

INTRODUCTION

The Personnel and Training Evaluation Program (PTEP) is the evaluation component within the SWS Training Program (described in Part I). By providing feedback to all of the other training system components, PTEP makes the SWS Training Program a self-correcting system. PTEP procedures, structure, organization and information handling systems were developed over a number of years and are constantly being improved to meet and carry out the evaluation and feedback responsibilities. These responsibilities include assessments of training facilities, hardware, documentation, individual courses of instruction and overall knowledge and skill of crew members. PTEP collects and analyzes data from many sources; provides reports; and, when appropriate, makes recommendations to those responsible for individual crews, schools, instructors and curriculum developers.

An effective evaluation and feedback system is essential to the development and maintenance of the SWS Training Program. PTEP evaluation is a continuous process examining various aspects of the training system as data bases and personnel change. New problems arise. New training courses are developed. Current curricula is revised. Instructors and instructional staffs are rotated. Characteristics of personnel entering the training pipeline change. New classes of SSBNs are designed and built. Current SSBNs are backfitted with newer systems, and other SSBNs are retired. PTEP has a very important charter which equates to a difficult and demanding set of responsibilities in a dynamic and ever changing environment.

The evaluation methodologies, statistical analyses and basic testing philosophies for the SWS Training Program have remained relatively constant. However, the changing environment and the multiple requirements lead to a competition for limited evaluation resources. And, the dynamics of this PTEP evaluation environment result in frequent realignments of project priorities. Therefore, managing the PTEP evaluation efforts, coordinating separate but related assessments, and communicating between and among the primary responsible organizations and agencies became a very complex and demanding responsibility in and of itself. Hence, the focus of this paper--to communicate how those responsible for SWS PTEP have made it easier to do a better job while keeping responsible parties informed.

BASIC ENVIRONMENT LEADING TO FORMATION OF AN EXECUTIVE STEERING GROUP

The SWS Training Program was (and is) so dynamic that control, coordination and communication of evaluations in this environment needed immediate and increasing management attention. A suggestion to form an executive steering group was put forth, accepted and the first meeting was planned in January 1981. An agenda was developed, distributed, and expected attendees were notified.

There are three primary participants engaged in the day-to-day operations of PTEP evaluation projects (The U.S. Navy Strategic System Program Office, Central Test Site and detachments, and contractor personnel). These participants have, as one might expect, different chains of command. They are geographically dispersed. PTEP evaluation responsibilities continued to encompass personnel, the training system, hardware, management documentation and facilities. In this environment one should not have been surprised to find that routine control, coordination, and communication methods would have been strained to the limit. They were. And in fact, there was a growing sense of frustration and even some doubt as to whether all of the evaluation responsibilities assigned to PTEP could be managed effectively.

It should be pointed out that the SWS Training Program takes evaluation very seriously, is committed to evaluation as an intergral part of the system, and relies upon a wide range of evaluations to help make management decisions regarding the training and preparation of our strategic submarine force. For example, during a five year period between 1976 and 1980 the SWS Training Program conducted eighty-eight studies that produced 428 recommendations related to most components of the training program, its operation, and the evaluation thereof.

It is important for one to realize the magnitude of the training program to understand the complexity of the PTEP evaluation requirements. As of July 1981 the SWS Training Program had a total of 180 courses. Ninety-five of these courses had completed all five developmental phases and had been promulgated. The content and conduct of these courses are considered stable. The eighty-five remaining courses were in various stages of development. Eight were in Stage 1 (scheduling planning), sixteen in Stage 3 (sample curriculum), forty-four in Stage 4 (pilot course), and sixteen in Stage 5 (Revisions, final course). These conditions make the actual conduct of evaluation and particularly the interpretation of the results even more difficult.

Another measure of the magnitude of PTEP evaluation responsibilities can be seen from another set of data. Approximately 126,000 records were created during a recent one-year period. During this same time 2,856 groups (average of 7.3 examinees per group) were tested--20,849 examinees' tests were scored, data recorded and results analyzed. All of this testing generated 85,828 original pages of reports with 177,432 additional pages reduced and photocopied. The primary PTEP data files used, at that time, consisted of 106.79 million characters of data storage.

PTEP evaluation responsibilities include personnel, training system, hardware and facilities. The testing of personnel and the scoring and analysis of the data represent a major undertaking. The magnitude of this evaluation responsibility is characterized in many ways. New tests must be designed and developed to replace current versions. Attention must be given to test reliability and validity of the data collected as curricula change. Another aspect of the changing character of this environment is embodied in the additions to SSBN's crews and the losses due to orders ashore, or completion of an enlistment and leaving the Service. Modernizing equipment adds still another dimension to the evaluation environment as older classes of boats are retired, newer classes are backfitted and still new ones built.

This is the real world of PTEP evaluation. It is dynamic, a constant challenge, and one that requires continuing diligence to all aspects of the process. One cannot forget the need for rigorous curriculum development; the classroom and laboratory exercises; the test design, development and administration; the analysis of the large quantities of data; and the presentation and interpretation of the results to those who need to know.

A few months ago it became evident that the routine methods of control, coordination, and communication within the PTEP component of the SWS Training Program were not functioning effectively. The initiation of an Executive Steering Group was an attempt to gain more control and more effectively manage the variety of PTEP evaluation projects. We wanted to deal with the more important issues and get out of the business of putting out brush fires.

INITIATION OF THE EXECUTIVE STEERING GROUP

In this dynamic pressure-packed environment we convened our first ESG in January 1981. Again from hindsight, it is not surprising to report that much of the first meeting was devoted to communicating between and among ourselves as to the current status of projects, perceptions, and problems facing us. Several action items were identified and responsibility for them was assigned during this one-day meeting. The group agreed that the first meeting had been worthwhile, and that we should meet about once a month for the next few months. Subsequently, minutes of the meeting were prepared and distributed to the attendees.

Later, an agenda was developed and distributed for the February meeting. Again the group received/gave status reports, adjusted schedules for projects, and determined action items. At this meeting the group also heard and considered some very serious internal criticism. The ESG found that by providing an opportunity to raise some very frustrating and potentially counter productive issues and perceptions and by dealing with them in a constructive manner, we were maturing as a group. The response of the ESG (in session), and actions by individuals as a result of the discussions constituted a significant step in the development of a genuine "team" spirit.

Development of an agenda, subsequent preparation of minutes, providing status reports, and realigning of priorities and schedules became routine. ESG meetings were conducted in January, February, March and April.

The next major milestone in the development of the ESG "management team" involved the use of a computer to keep track of the growing number of action items. During May we developed a relatively simple automated system for tracking our action items. Each action item was (and is) assigned a six-digit file number. The first two digits equal the last two digits of the calendar year, the second two digits equal the month of the ESG meeting, and the last two digits equal the item number used to record the essence of topics discussed in the ESG meetings.

This relatively simple file-number system permits one to rather quickly refer to the appropriate minutes for more details, when desired. The automated action item file contains, in addition to the file number, a description of the action item (200 character limit), due date, complete date, primary responsibility code, support responsibility code, and a remarks field (200 character limit). Listings of action items are provided ESG members at various times for use in managing items assigned and to report completion or a change in the status.

Prior to the June ESG meeting each member was furnished with a complete listing of all action items (62) since the January meeting. At the June meeting the purpose and uses of the action item listings were discussed. Attendees were asked to provide an update for each item for which they had primary responsibility.

Following the June meeting a new listing was provided for each primary responsibility code and attached to the ESG minutes. This new listing included the action items from the June meeting.

Another milestone in the development of the "management team" was reached in July when the ESG adopted a new format for conducting subsequent meetings. Essentially, the new format simplified the ESG agenda. Meetings will be conducted by first reviewing and approving the previous minutes with any corrections. Second, the ESG will review the open action items from previous meetings. And lastly, new topics and issues will be presented and discussed. The new format simplifies the agenda, focuses specific attention on current status of assignments and makes it easier for participants to recognize new business.

A complete set of ESG minutes and an action item listing were a great source of information for a recent ESG member who joined the group for the first time in July 1981. These two sets of documents summarized for him what had been accomplished and discussed over the past few months, the time sequence in which decisions had been made, and the relationships of certain action items. Some items are cross-referenced by file-numbers in the remarks data field. As other ESG participants join this group, we will see even more uses for the effort invested in resolving important issues as a group, recording the results of these actions, in tracking and maintaining control over the one-time and continuing evaluation activities.

Another milestone in the development of the "management team" was reached in the August meeting. This was the first time that representatives from Central Test Site detachments had attended these meetings. The detachments are very much in the front line of PTEP test administration and of feeding back the results to the SSBNs. Very often they (detachments) are the first to sense or hear of a problem, and it is important that they understand why certain projects are being undertaken and how they are progressing. Additionally, a representative from Chief, Navy Technical Training (CNTT), also sat in on the meeting. The Central Test Site is under the command of CNTT. Both the detachments and CNTT representatives had an opportunity to see how the ESG operates and the full range of PTEP evaluation responsibilities being controlled and coordinated by this "management team." The openness of the ESG to input from and observation by others is another indication that the group has experienced some growth.

The ESG has found a balance between managing the evaluation process and in utilizing the technical expertise necessary for the development of evaluation methodologies, implementation of these plans, and in the conduct of evaluations, per se. Technical expertise abounds within the resources available to the ESG. There are subject matter experts familiar with the complete range of curricula, hardware and documentation.

Other members are well trained in statistical analysis, psychometrics, educational research, test design and construction, information systems, and management science (including organization development). Individual commitment to PTEP evaluation by this very diverse group also made a contribution to the limited success of the ESG. Some members contribute more from a technical point of view while others are more attuned to the process by which we collectively seek to achieve long term dividends from all PTEP evaluation efforts.

CONCLUSIONS AND SUMMARY

The ESG came into being (was born) out of a genuine need. The environment was ripe for something which would provide more control, integrate and coordinate all of the evaluation efforts, and at the same time improve communications. It took a great deal of patience, attention to detail, emphasis upon preparation and following up of meetings, and a diligence toward the organizational development process. In the beginning, some of the ESG members were more open and trusting than others. Each individual has a separate set of perspectives and brings them to the ESG. And we found value in and some truth to each of these perspectives. In our attempt to gain better control, move the PTEP evaluations forward, and to grow as professionals we were forced to learn from each other and about ourselves. We recognized anew that often growth does not come easy. It is sometimes hard to listen, and even more difficult to really hear that which one had rather not be told. However, we have listened and have learned from each other. The author feels that the members of the ESG experienced a great deal of growth during the past several months as we dealt with evaluation issues from many different perceptions.

As a group, we undertook several new tasks designed to improve the overall performance of PTEP. These actions were carried out in the dynamic environment described earlier. Schedules were adjusted, priorities were realigned, and resources were reallocated in light of new information. The ESG did not try to resolve all difficulties, but did try to deal with differences in a mature and understanding manner. We attempted to focus upon tasks to be accomplished, while growing in our capacity to accept the fact that we could disagree without being altogether disagreeable. We have achieved a measure of success on both counts. A review of the major milestones confirms this fact. Also, the number of action items assigned (100) and completed (82) during this period also attests to the fact that we had gained better control of evaluation projects than one would normally expect with conventional management methods. The underlying commitment of individual ESG members to the process of evaluation as embodied in PTEP for the SWS Training Program helped us band together and to develop this "management team."

The importance of the ESG in strengthening PTEP evaluation efforts resulted from a lot of hard work. The results described above did not suddenly and miraculously burst forth in response to our real need. Much of the work went on between the meetings themselves that was supportive in nature, that took care of important details, and was sensitive to good management principles and organization development The author does not intend to suggest that ESGs are "the answer" to other evaluation efforts, but rather that this process has served our group well.

In this context a few suggestions are offered to those who might want to use the basics of this concept in another environment. These suggestions are grouped into three categories. First, keep things as simple and straightforward as possible. Make good use of time by developing and distributing agendas before the meetings. Then conduct the meetings as "working sessions" in accordance with the agenda. Keep the team small and productive -- not a convention. Subsequent to each meeting prepare and distribute brief minutes that are reviewed, corrected and approved at the next meeting. Second, the author would suggest that leadership of the group should recognize, accept and be able to deal with different perspectives. The author means by this statement that an opportunity should be given for different points of view to be expressed before important decisions are made that will have an impact upon the group. Likewise, it is important for this leadership to follow up on differences that exist between and among members so that they can continue to express differences without being altogether disagreeable. The third suggestion is that an audit trail or record of group decisions be provided. The minutes from the meetings may serve this purpose completely. If they do not, then other measures should be taken. For example, when the total number of actions is quite large, membership in the group changes, and there is a need to retrieve selected material from time to time, an automated system may be required in addition to the minutes. The author also considers the audit trail an important tool or device for orienting new team members as well as those who have an occasional need to know what is going on.

Evaluations are seldom conducted where all of the variables are well controlled. In PTEP, only a few of the variables are controlled. Hence, the analysis and interpretation of PTEP data are difficult, but not impossible. The key is in creatively using what is available and in managing the process well. The author is pleased with the role that the ESG has played in PTEP evaluation and anticipates that this "management team" will continue to be an important means by which we maintain control, provide appropriate levels of coordination, and supplement communications between and among the key participants in the dynamic SWS Training Program.

Effect of Expectations on Attrition in Canadian Forces Recruit Training

Captain D. S. Horton
Base Personnel Selection Officer
Canadian Forces Base Gagetown
New Brunswick, Canada

This study investigated the effect of expectations on attrition in Canadian Forces (CF) recruit training. The 820 male participants, ~~included~~ 699 who successfully completed recruit training (stayers) and 121 who failed or withdrew (leavers). Participants completed a pre-test questionnaire within two days of arrival at recruit school and a post-test questionnaire after three weeks of training. Expectations of both groups were measured against a 12 member criterion group, comprised of senior recruit instructors, recruiting officers and personnel selection officers. Findings indicate that expectations of stayers and leavers were not significantly different nor did leavers report a greater degree of disconfirmed expectations. Leavers did, however, have a greater ratio of expectations disconfirmed in a "worst than expected" direction. A relationship was found between attrition and expectations of success in recruit training, membership in the CF after recruit taining and intended length of service, with leavers responding more negatively in all three areas. Although assessment of pre-training information did not differ for the two groups, the responses of both groups shifted in a negative direction on the post-test. Implications for the CF, of the findings, are discussed.

As inflation continues to escalate, most segments of western society are experiencing increased financial pressures. This is particularly so in military organizations, where it is becoming increasingly difficult to justify huge monetary expenditures to adequately train and equip military forces. The Canadian Forces (CF) has, in recent years, been closely scrutinized by both government and citizens in regard to monetary expenditures, and it has become imperative that every dollar of the military budget is spent wisely and effectively. This close monetary scrutiny, while affecting all aspects of the CF, has been particularly focused on the recruiting and selection processes utilized to maintain necessary manpower levels. Recruitment, enrolment and retention of suitable personnel has become an increasingly critical issue.

In addition to imposed financial constraints, recent research (Cotton, 1974) indicates that the pool of potential recruits for the CF will begin to shrink in the early 1980s. Concurrently, competition with industry for the available manpower pool is increasing steadily. As well, the CF will require a greater proportion of high quality personnel to cope with advancing military technology. Financial constraints, increased competition for recruits and the need for better recruits, demand that recruiting becomes increasingly cost effective, with attrition at the recruit training level becoming a critical problem.

It is doubtful that attrition in recruit training is precipitated by a single definable factor, but more likely results from a number of interacting factors. The factor of expectations is considered important in explaining attrition from any organization. Extensive research has supported the assumption that prospective organization members bring with them a complex expectation set of their role and membership in their chosen organization. Wiskoff (1977; Page V.) observed that "expectations have been shown to exercise a pervasive influence on decision processes of individuals to join, develop occupational choices, remain, or derive satisfaction from the military." That the expectations brought by prospective organization members are often realistic has also been well documented by previous research. This phenomenon may be particularly true for new members of the military who encounter an environment that is alien to previous experiences.

In considering the relationship between expectations and attrition, the degree, area and direction of disconfirmation of expectations are of primary interest. Previous research indicates it is not necessarily the degree of disconfirmation that contributes to attrition, but the direction or area of disconfirmation, or a combination of both direction and area. This study investigated the interrelationships of degree, direction and area of disconfirmation of expectations and their relationship with attrition in CF recruit training. Although previous CF studies have acknowledged that expectations are a major contributing factor to attrition during recruit training [eg., Cotton (1974, 1975); Fournier and Keates (1975); Rampton (1973); Mullin (1977)], there has not been a sustained focus on the extent to which expectations effect attrition. The implications of research in this area could be significant for the CF, with the possibility of reduced attrition during recruit training, a concomitant reduction in the unsettling effect leavers have on other recruits and, reduced recruiting quotas resulting from reduced attrition.

In view of continuing financial constraints, a shrinking recruit pool, increased competition with industry for recruits and high attrition rates during CF recruit training (averaging 20% between 1974-1979), it is timely to examine the effect of expectations on attrition during CF recruit training. This study attempts to answer a number of critical questions regarding this phenomenon. To begin with, are the expectations of recruit training held by leavers less realistic than those of stayers? Do leavers report a greater degree of disconfirmed expectations? If so, are more leavers' expectations disconfirmed in critical areas and/or in a worse-than-expected direction? Do leavers have a less positive expectation set toward membership in the CF after recruit training? Is the expectation of success in recruit training less for leavers? Is the intended period of service less for leavers? Do leavers classify pre-enrolment information as being less accurate and sufficient than do stayers?

Previous Research

The various theoretical models developed in recent years to explain the influence of expectations in human behavior offer interesting implications for prospective or new organization members. Current theory suggests a strong relationship between expected and actual outcomes, and their attractiveness to the individual, and the energy expended toward becoming a productive organization member. Research supports the contention that the act of joining a group or organization is carried out the for purpose(s) of goal and need satisfaction. Upon committing himself to an organization, the individual has, as much as possible, become convinced through assimilated information from various sources, that expected goals and needs can be attained. Gauntner (1973), Owens (1970), Weston (1974) and Youngberg (1963) all alluded to the strong relationship between pre-enlistment expectations and satisfaction with the military of new members. Rampton (1973) commented that if an organization hopes to "attract or keep sufficient personnel" then it must be perceived "to offer an environment that meets the individual's expectations and needs as well or better than other available options" (p. 28). The pervasive influence of expectations in determining job motivation, performance, satisfaction and turnover has been well documented by research. Macedonia (1969), Owens (1970), Mullin (1977), Wiskoff (1977), Bourne (1965), Griffith et al (1979), Ilgen (1975), Scott (1972) and Youngberg (1963), all commented on the potential for dissatisfaction and eventual attrition if expectations of what will happen in a stressful environment, such as recruit training, are unrealistic and therefore unmet.

For the military, the group most strongly identified with the question of realistic and/or unrealistic expectations of new members are the recruiters. Expectancy theorists have been less than kind to military recruiters, openly questioning their integrity and credibility. As the interface between the military and the potential enrollee, the recruiter faces a difficult challenge in presenting one to the other. This is particularly true in the process of developing, in the applicant, a realistic expectation set of recruit training, which inevitably includes modifying unrealistic expectations. In fairness to recruiters, the majority probably do it as best they can; some may not appreciate the critical role they play in shaping short and long term attitudes of military recruits.

Basically, the recruiter's role is to close the expectation gap between the recruit and the organization. In recent years, extensive research has been conducted in seeking effective methods of closing the expectation gap. Much of it has focused on the utility of providing realistic and accurate information to new organization members. Weitz's (1956) study was a landmark in this area when he demonstrated that provision of realistic information resulted in reduced attrition and an increased tendency to join an organization. Macedonia (1969) produced similar results in a study of entering USMA cadets. He found that provision of realistic information ferreted out impulsive applicants, that enrollees accepted greater personal responsibility for their decision to join and that enrollees were better prepared for the situational stresses with which they were confronted. Studies by Farr et al (1972), Bittinger (1964), Wanous (1973) and Sherman (1959), produced similar results. Scott (1972) emphasized the necessity for recruiters to realize that selection is a mutual process involving both the applicant and the organization. Scott believed that selection is too frequently approached from the organization perspective only, with recruiters gathering reams of information about the applicant but not reciprocating with necessary information about the organization. Wanous (1973) drew conclusions similar to Scott's, stating that industrial psychologists are pre-occupied with the organization's viewpoint, at the applicant's expense. The 1973 study by Wanous utilized the concept of "realistic job previews", which resulted in more realistic job expectations and no difference in job acceptance rates between control and experimental groups.

Previous expectancy research has also focused on the problem of identifying potential leavers before or shortly after enrolment. Various theories have been postulated to resolve this issue. Ilgen (1975), Schuckit and Herman (1978) and Mullin (1977) believed that the impulsive individual, who enrolled without giving serious consideration to implications of joining the military, quite often fails or quits during recruit training. Expectations of success in early training and employment (eg. Mobley et al, 1977), level of formal education (eg. Lockman, 1976 and Cotton, 1974), category of service (O'Gorman, 1972) and age (Guinn, 1977), have also been put forth as indicators of potential attrition

Barrett (1975) in Wiskoff (1977) most succinctly summed up the large picture of what the far-sighted goal of all expectation theorists should be when he postulated that:

> . . . there is an optimal match or congruence among abilities, preferred attributes, expectancies and task complexities which will result in maximization of resources in terms of individual productivity, work satisfaction, and organization tenure (p. 20).

Method

Subjects: The study conducted in early 1979, consisted of 820 male recruits who underwent training at the Canadian Forces Recruit School. Of the 820 participants, 699 successfully completed recruit training (stayers), while 121 failed, or were released at their request (leavers). The heterogeneous sample included participants from every province and territory in Canada, a wide variety of socio-economic backgrounds, and participants ranged in age from 17 to 25 years.

Procedure: The study employed a six-part questionnaire for data collection. In Part I, demographic data were collected. Part II contained 55 items on which the respondent indicated how much of each item he expected to experience in recruit training. Parts III and IV contained the same items and format as Part I, with respondents being requested to indicate preferred amount of each item and importance of finding preferred amount respectively. Part V consisted of three items pertaining to (a) expected success in recruit training, (b) length of intended service in the CF and, (c) assessment of accuracy of pre-training information provided by recruiters. Part VI was a survey of post-CFRS expectations of membership in the CF. For this study, only Parts II, V and VI were analyzed. An evaluation of the questionnaire with a sample of 66 recruits led to minor modifications in format and clarification of some instructions. Item analyses of the questionnaire were conducted, and an alpha reliability coefficient of 0.94 was obtained. The instrument was designed by the author, with items being chosen on the basis of: recruit information films, recruiting brochures, consultation with CFRS staff and experienced recruiters and the author's own experience as a Personnel Selection Officer (PSO) in both recruiting and on operational military bases.

A pre-test questionnaire was administered during the recruits' first two days at the CFRS. The post-test was administered at the completion of three weeks of recruit training. Questionnaires were administered by the PSO staff at the CFRS to groups of participants averaging 83 in number with respondents being assured anonymity. The procedure resulted in 820 usable questionnaires.

The 55 expectation items were also administered to a 12-member criterion group which consisted of senior instructors and the PSO's from the CFRS, recruiting officers and the author. Arithmetic means of their responses to each item were established as an accurate measure of what recruits could realistically expect, in regard to the 55 expectation items, at the CFRS.

Results

Discriminant analysis procedures were conducted to compare responses of the two groups on both the pre and post tests. T-tests were also conducted using a $p < .05$ level of significance.

Expectations of leavers were not dissimilar from those of the stayers, in comparison to the criterion group. Leavers had 15 of the 55 expectations items significantly different from the criterion group while stayers had 16. Of the 15 items on which the leavers differed from the criterion group, 13 were the same as the items on which the stayers differed from the criterion group.

Disconfirmation: Leavers did not report a greater degree of disconfirmed expectations than did stayers. Leavers did have a greater ratio of expectations disconfirmed in a "worse than expected" direction. while disconfirmation for stayers was almost the same in both directions. Table 1 clarifies these results.

Table 1

Disconfirmed Expectations for Stayers and Leavers with Percentage and Direction of Disconfirmation

No. of expectations disconfirmed	Percentage Disconfirmed	"better than expected" direction	% better	"worse than expected" direction	% worse
Stayers:					
44	86.275	32	47.720	23	52.273
Leavers:					
30	58.824	9	30.000	21	70.000

Critical Dimensions: Eleven critical dimensions were identified using the discriminant analysis stepwise procedure. It was found that leavers did not report more disconfirmed expectations in critical areas than did stayers. Leavers did, however, have more critical dimensions disconfirmed in a worse than expected direction, than did stayers.

Post-CFRS Attitudes: Leavers were found to have a less positive expectation set about membership in the CF after recruit training. Table 2 illustrates these results: on the pre-test, stayers were less positive than leavers on seven items, while leavers were less positive on 13 items. On the post-test, stayers were less positive on three items while leavers were less positive on 17 items. This reveals a shift in Post-CFRS attitudes for both the stayers (more positive) and leavers (less positive) from the pre- to post-test.

Table 2

Post CFRS Expectation Sets for
Stayers and Leavers - Pre and Post Test

No. of items less positive (pre-test)	% items less positive (pre-test)	No. of items less positive (post-test)	% items less positive (post-test)
Stayers:			
07	35.0	03	15.0
Leavers:			
13	65.0	17	85.0

Expectations of Success: Leavers' expectations of success in recruit training were less positive than those of stayers, on both the pre and post-test. In fact, the negative shift by the leavers from the pre to post-test was also significant ($p < .05$). The leavers, therefore gave themselves even less chance of successfully completing recruit training, after three weeks at the CFRS. Table 3 demonstrates these results.

Table 3

Stayers and Leavers Expectations of
Success in Recruit Training - Pre and Post Test

	Stayers	Leavers	Significance of difference
Mean response			
pre-test	*4.9941	4.5576	.001
Mean response			
post-test	4.9842	4.0781	.001

*Range of response: 1 through 6.

Intended Length of Service: Leavers intended length of service was less than that of stayers. Stayers intended to serve longer, on both the pre and post tests, with their intention.

increasing from the pre to post test, albeit not significantly. Leavers intended length of service decreased from the pre to post test ($p < .05$). Table 4 elaborates on the results.

Table 4

Stayers and Leavers Intended
Length of Service - Pre and Post Test

	Stayers	Leavers	Significance of difference
Mean response			
pre-test	*7.1073	6.6552	.006
Mean response			
post-test	7.1848	6.0500	.001

*Range of response: 1 through 8.

Assessment of Pre-Training Information: No difference was found between stayers and leavers in assessment of pre-training information. What is particularly interesting about this hypothesis, however, is the dramatic difference ($p < .05$) between pre and post test responses for both the stayers and leavers. On the pre-test 71.45% of stayers and 67.36% of the leavers assessed pre-training information as being accurate and sufficient. On the post test only 37.61% of the stayers and 32.26% of the leavers assessed pre-training information as being adequate and sufficient. Table 5 illustrates these results.

Table 5

Stayers and Leavers Assessment of Pre-Training Information - Pre and Post-Test

	Stayers	Leavers	Significance of Difference
Mean response			
pre-test	.7145	.6786	.468
Mean response			
post-test	.3761	.3226	.400

Discussion

Findings regarding expectations of CF recruit training were not consistent with other recent studies indicating a difference in expectations between stayers and leavers. The fact that this study did not reveal a difference in expectations of recruit training between stayers and leavers suggests that accuracy of expectations, in general, may not be a decisive factor in influencing attrition in CF recruit training.

In this study, degree of disconfirmation of expectations did not differ significantly for stayers and leavers. These findings also differ from recent research and suggest that degree of disconfirmation is not a decisive factor in attrition of CF recruits. Owens (1970), reached the same conclusion, based partly on the assumption, also documented by other studies, that new organization members (particularly military) lack previous experience on which to draw in developing a base of realistic expectations. Adjustment to disconfirmation of expectations, therefore, appears to be a salient factor in explaining attrition during recruit training, with previous studies [Wiskoff (1977); Wadsworth (1974); Hammond (1966)] indicating that adjustment was usually in a downward direction.

Direction of disconfirmation also appears to be a differentiating factor in attrition, with leavers experiencing more disconfirmed expectations in a "worse than expected" direction. In particular, disconfirmation in a negative direction in critical areas (eg: specifically identified variables) may be decisive factors in influencing attrition.

In a 1977 study, Mobley, Hand and Logan attributed the differences between stayers and leavers ability to adjust to disconfirmation as being due to differences in "certain personality trait dimensions" such as confidence, anxiety, self-image, etc. These perceived traits negatively affected some individuals' ability to cope with the anticipation and reality of heavy stress (leavers), while others were more flexible in adjusting expectations to reality.

Pre-enrolment identification of potential leavers appears feasible from results of this study. It was found that leavers have a less positive attitude than stayers toward membership in the CF after recruit training, give themselves less chance of successfully completing recruit training, and are more tentative in thier commitment to the CF. These results support Wiskoff's (1977) contention that "expectations about the future exercise more influence as a determiner of career intent than present conditions" (p. 27). Porter and Steers (1973) also saw expectations of the future as having a pervasive influence on career intent and career decisions of recruits. Identifying potential recruits who are negative or doubtful about what to expect in the military would facilitate either screening-out procedures or, provision of counselling at the selection stage to modify attitudes and expectations.

The fact that both stayers and leavers were similar in their assessment of pre-enrolment information was somewhat surprising. Previous research indicated that leavers' assessment of such information would be less positive than stayers. The fact that almost one-third of both groups were dissatisfied with pre-enrolment information, even before commencement of recruit training, suggests several possibilities. It could be that information being given recruits about the recruit training experience is in fact not accurate and sufficient or, for some reason the recruits are not internalizing the information. Possibly, a combination of the above may occur, partly due to the phenomenon labelled "selective hearing". The significant negative shift in assessment of pre-training information after three weeks of training warrants serious consideration. If only the leavers' assessment had shifted negatively, one might attribute it to rationalization and displacement of responsibility for the inability to adjust to the rigors of recruit training. These results do suggest that when disconfirmation does occur, it is attributed to pre-training information. An examination of pre-training information may be in order. These findings also further support the contention that stayers are better equipped to adjust to disconfirmation.

Reinharth and Wahba (1975) postulated that the expectancy model could emerge as an influential approach to explaining work behaviour if it could be proven to effectively "describe and predict work motivation, job effort and job performance" (p. 521). Utilization of the expectancy model to effectively discriminate between stayers and leavers in CF recruit training lends support to their postulation. The implications of using the expectancy model to identify potential leavers before or shortly after enrolment could be far-reaching. Measurable savings in human and financial resources could accrue from the ability to differentiate between these two groups during the enrolment process. In these days of a shrinking recruit pool, increased competition from industry for recruits, and continual financial constraints, the ability to reduce attrition is crucial. Analysis of the remaining data collected in this study should provide further insight into the role of expectations in attrition during recruit training.

References

Barrett, G.V. Task design, individual attributes, work satisfaction, and productivity, Abron, Ohio: University of Abron, 1967. (Presented to Conference on the Current Studies on Work Organization, Berlin, West Germany, October 13-15, 1976).

Bittinger, E.F. Realism and Vocational Expectations: Increasing Realism with Advancing Age, University of Maryland, 1964.

Bourne, P.G. Some Observations on the Psychosocial Phenomenon Seen in Basic Training, Department of Psychiatry, Walter Reed Army Institute of Research, Washington, D.C. 1965, 187-196.

Cotton, C.A.U. Research on socio-demographic factors which influence military service in Canada (Working Paper 75-5). Canadian Forces Personnel Applied Research Unit, August, 1975.

Cotton, C.A.U. Unskilled Male Recruits For the Canadian Forces: Trends and Projections to 1985, Canadian Forces Personnel Applied Research Unit, Research Report 74-5, Toronto, 1974.

Farr, L.J., OLeary, B.S., Bartlett, C.J. Effect of Work Sample Test Upon Self-Selection and Turnover of Job Applicants, Journal of Applied Psychology, 1972, 38, No. 12, 283-285.

Fournier, B.A., Keates, W.E. Why Do They Leave: Attrition, 1975, Canadian Forces Personnel Applied Research Unit Report 75-4, September, 1975, Toronto.

Gauntner, D.E. An Analysis of Volunteer Soldiers' Perceptions and the Expectations About Satisfaction with Army Life U.S. Army War College, Carlisle Barracks, Penn., May, 1977.

Hammond, D.A. Observations of The Areas of Army Service Seen as Attractive by Serving Soldiers, Research Report 4/66, Australian Army Psychological Research Unit, Melbourne, October, 1966.

Ilgen, D.R. The Influence of Expectations and Beliefs on the Motivation and Adjustment of New Members of Military Organizations. Presentation to a Research Seminar on Social Psychology of Military Service, 23-25 April, 1975. Center for Continuing Education, University of Chicago.

Lewin, K. A Dynamic Theory of Personality, New York: McGraw-Hill, 1935.

Lockman, R.F. Forecasting Enlisted Attrition: The First Year of Service, Centre for Naval Analysis, 1975.

Macedonia, R.M. Expectations-Press and Survival, Unpublished PhD dissertation, New York University, 1969.

Mobley, W.H., Hand, H.H., Logan, J.E., Baker, R. Pre-Recruit Training Values, Expectations and Intentions of Marine Corps Recruits. Center for Management and Organizational Research, University of South Carolina, May, 1977.

Mullin, E.E.D. Project Profile: An Analysis of Male Recruit Attrition, Canadian Forces Personnel Applied Research Unit. Research Report 77-6, Toronto, 1977.

O'Gorman, J.G. Review of Research on Re-engagement Among Australian Regular Army Other Ranks. Research Report 2/72: Australian Army Psychological Research Unit, Melbourne, 1972.

Owens, A.G. The Study of Assimilation into the Armed Services. Paper delivered to the meeting of UTP-1 of TTCP, 29 April - 1 May 1975. Psychology Research Unit, Chancery House, Melbourne, Australia.

Owens, E. The Effect of Disconfirmed Expectations on Adjustment to Cadet Life at The United States Air Force Academy. Unpublished PhD dissertation, University of California, 1970.

Porter, L.W., Steers, R.M. Organizational, work, and personal factors in employee turnover and absenteeism. Psychological Bulletin, 1973, 80, 151-176.

Rampton, G.M. Recruiting Shortfalls and Personnel Policies Within the Canadian Context. Canadian Forces Personnel Applied Research Unit. Nov. 1973.

Reinharth, L., Wahba, M.A. Expectancy theory as a predictor of work motivation, effort expenditure, and job performance, Academy of Management Journal. 1975, 18 520-537.

Schuckit, M.A., Herman, G. Pre-Enlistment Decision Time as a Predictor of Military Effectiveness - Alcoholism and Drug Abuse Institute, University of Washington, Seattle, 1978, 171-174.

Scott, R.D. Job expectancy - an important factor in labor turnover. Personnel Journal. May, 1972, 360-363.

Sherman, E.L. Turnover cut one-third by seven-step program. Personnel Journal. Jan. 1959, 296-297.

Tolman, E.C. Purposive Behavior in Animals and Men. New York: Appleton-Century, 1932.

Vroom, V.H. Work and Motivation, New York: Wiley, 1964.

Wadsworth, E.L. The Relation of Met and Unmet Expectations to Subsequent Performance by Trainees in the Job Corps YWCA Extension Residence Program. Unpublished PhD dissertation, New York University, 1974.

Wanous, J.P. Effects of realistic job preview on job acceptance, job attitudes and job survival. Journal of Applied Psychology, 1973, 58, 327-332.

Weitz, J. Job Expectancy and Survival. Journal of Applied Psychology, 1956, 40, 245-247.

Weitz, J., Nuckols, R.C. Job Satisfaction and Job Survival. Journal of Applied Psychology, 1956, 39, 294-300.

Weston, P.J. A Person-Environment Analysis of Performance, Satisfaction and Expectancies of Job Trainees. Unpublished PhD dissertation, University of Michigan, 1974.

Wiskoff, M.F. Review of Career Expectations Research: Australia, Canada, United Kingdom, and United States. Navy Personnel Research and Development Center, San Diego, California, March, 1977.

Youngberg, C.F. An Experimental Study of Job Satisfaction and Turnover in Relation to Job Expectations and Self Expectations. New York University, 1963.

Hosmer, Clark L., COL, USAF (Ret.), Shalimar, Florida. (Thurs. P.M.)

The Honor Code: The Optimum Criterion of Honor?

The Honor System at West Point is renowned as the most important single influence on the lives of graduates. A cadet must be honest and, harder yet, not tolerate his friend's dishonesty. Despite the difficulties of rectitude plus being his brother's keeper, only a few cadets violate the system in most years. A major problem is failure of the non-toleration policy associated with outbreaks of group cheating - six outbreaks in thirty years. Four programs to improve implementation of the non-toleration policy are offered. If future evidence shows continued failure of the non-toleration policy, an alternative criterion is offered for consideration.

The Honor Code: The Optimum Criterion of Honor?

Clark L. Hosmer
39 Longwood Dr.
Shalimar, Fl 32579

My purpose is to assess the Honor Code as a criterion of the Honor System at the United States Military Academy at West Point, New York. The Honor Code is, "A cadet will not lie, cheat, or steal nor tolerate those who do."

The Honor Code serves as, "... the minimum standard of behavior required by cadets. Because of the Code, the Corps enjoys an atmosphere of trust and the reputation of being honorable. The Code is also a foundation for continual development of a cadet's personal standards and values." (U.S.M.A., 1979).

The Honor System is not 100% successful in converting all cadets to honorable behavior. Usually fewer than one percent of the cadets are found guilty of violating the Honor Code in a year. Cadets come from all walks of life in the nation where six out of ten teen-agers admit cheating on examinations (Gallup, 1979). Some cadets are unable to make the transition to exemplary honorable behavior. Not only must the cadet be honest but the Honor Code demands that the cadet not tolerate another cadet's violation of honor. The cadet is expected to be as offended by a friend's violation of honor as by his or her own violation. Woe to both the violator and the tolerator. Both are subject to dismissal from the academy. "He that worketh deceit shall not dwell within my house... nor tarry in my sight." (Psalms 101:7).

Except for the few violators, the cadets who survive the highly competitive regime of academic studies and military training develop a camaraderie of trust among themselves. They become imbued with the rigor of the Honor System that demands rectitude despite discomfort or cost to self or others. The Honor Code becomes a shared moral imperative and criterion for service described by West Point's motto, "Duty Honor Country." Graduates often describe the Honor Code as the single most powerful, enduring, and beneficial influence that the academy has had on their lives.

<u>Vulnerability to Group Cheating</u>

An outbreak of group cheating has occurred at West Point six times in the latest thirty years. In 1976, the group involved more than [illegible] cadets. The episode was of such magnitude and complexity that the Secretary of the Army appointed a special commission to study, "... underlying causes in the context of the Honor Code and Honor System and their place in the Military Academy." (Borman, p. 5).

Official reports cited toleration of honor violations as a major problem for four years before the 1976 outbreak of group cheating. Then in 1974 a survey of the cadets showed that 73% would not report a good friend for a possible honor violation and 34% would not report a good friend for a clear-cut violation. Forty-five percent of the cadets wanted toleration removed as a personal honor violation (Borman, p. 14). Although toleration was an obvious problem, for a period of ten years before 1976, conviction of cadets for toleration alone was only one-half of one percent of all convictions of honor violations (Borman, p. 17). That finding suggests that toleration of toleration was almost the rule.

Toleration of another's honor violation creates vulnerability of the Corps of Cadets to group cheating. Before the 1976 episode, the Superintendent's Honor Review Committee reported, "... any cheating scandal would find its beginning in a 'toleration' situation, i.e., a cadet would observe a friend or roommate cheating but because of their closeness would not report the incident. From that point a vicious chain would gradually find its way to other cadets." (Borman, p. 7).

The evidence cited in the Borman report suggests that the Honor Code may not be the optimum criterion for the Honor System. The non-toleration clause was ineffective which opened the door to possible group cheating.

To Improve the Effectiveness of the Non-Toleration Clause

The four following programs may further the progress already made by the Corps of Cadets and the academy's superintendent and staff and faculty toward improving the Honor System. The objective of the programs is to reduce the rate of cadet readiness to tolerate another's honor violation and thereby reduce vulnerability of the system to group cheating.

1. A program to collect anonymous responses of a random sample of cadets to well designed items that are repeated every quarter plus special items on problems that come up would provide valuable objective data. Significant trends in strengthening of, or trouble in, attitudes toward non-toleration would be on the table for handling. Another source of data could be coverage of the non-toleration policy in the exit interview of every cadet who leaves the academy whether for academic deficiency or honor violation or other reason. Due to the freedom the departing cadet has to ventilate privately held views, the exit interview would collect information not anticipated by the regular items used in the quarterly surveys. Strong views associated with reasons for departure could be assessed for their influence on answers to the regular items. A third source of information could be invited papers by cadets in professional ethics and leadership classes. If each cadet were invited each year to submit a paper on any aspect of the Honor System that he or she selects, the range of system strengths and problems, let alone suggestions, would provide trends in elements faring well and those that

warrant special attention of the Cadet Honor Committees and perhaps the academy administration as well. Four thousand papers would be mountainous, of course, but sorting of topics and harvesting of suggestions on pivotal subjects like non-toleration would, in my view, justify the efforts by cadets and analysts.

2. A program of demonstration by staged skits would help to exercise the new cadets in handling of the problems of identifying a violation, counseling with the violator to make sure of the facts, and seeing that one or both report the event to an Honor Representative. Skits based on historical cases without identification of persons, would heighten realism. Role-playing the world of cadet life, to include upperclassman roles, would exercise upperclassmen, too, in handling the classic tough case of seeing a close friend, or for many the even tougher case of seeing a supervisor, violate the honor code. The skill required to react promptly with tact and yet with effective force does not come from exhortation to do so. Skits have unusual powers in terms of illustrating a range of problems and developing confidence to confront rare events.

3. A program of new and old cadets in/different small groups, with a minimum of direction by a supervisory level chairman, could provide self-convincin opportunities for understanding the imperativeness of the non-toleration policy. The most effective way for people to understand and accept new ideas is in free-for-all discussion of all the pros and cons. Most of the groups of people who were allowed to talk freely among themselves about the pros and cons of a system chose loyalty to group interests. Without complete freedom of discussion in confronting a new system, however, 75% of small groups chose to strive for individual interests. The individual competition reduced gains for individuals as well as for the groups as a whole (Edney, p. 84).

4. A program to publicize every quarter selected findings from the information garnered by the activities above could keep the Corps of Cadets focused on implementation of the non-toleration policy. The suggestion of publicity is controversial. If a survey shows a third of the cadets are skeptical about non-toleration, would reporting that finding to all the cadets suggest to some of them that they may as well join the tolerators? Whatever the level of skepticism, to share that information within a program aimed at lowering vulnerability to group cheating, in my view would pay in the long run. Would leakage to the public media about a third of the cadets being skeptical hurt the Corps? The national furor about the 1976 cheating scandal was a blow but media reactions included staunch defenses of the splendid reputation for honor in the Corps that soon again the solid reputation would be warranted. The problem of vulnerability to group cheating is too important, in my view, to use usual staffing of solutions and announcement of what is to be done. The Honor System is the cadets' as backed up by the superintendent and his staff and faculty. Unless the cadets accept the system including the non-toleration policy, vulnerability to group cheating

will continue. The best bet, in my view, for keeping vulnerability at a minimum includes the cadets themselves **knowing** the objective score. For the most part they will solve the problems themselves. Exhortation will not solve the problem. Concerted internal Corps programs would seem to be the most effective approach.

Is an Alternative Criterion Available?

There is an old saying, "If it ain't broke, don't try to fix it." If, however, future evidence shows that the non-toleration policy is fully accepted by only, say, 75% of the cadets, would consideration of an alternative criterion for the Honor Code be warranted? In case objective evidence of cadet determination to implement the non-toleration policy is less than the level that the Corps sets to prevent group cheating, I submit a possible alternative Honor Code for consideration.

On its establishment in 1802, West Point adopted a code, the core of which was, "A cadet is fundamentally honest and therefore accepted at his word." (U.S.M.A., 1977, p. 6). Those words show the compelling reason for, and the result of, being honest. Also, the statement is positive. In my view the affirmative has power that prohibition of lying, cheating, and stealing does not match. "Fundamentally honest," is broad-brush as contrasted with the specificity of the prohibitions in the present Code. Honor itself lacks specificity. The original Code may be more enlightening for its scope than the short inventory of three evils featured today among the many such as bribery, fraud, forgery, plagiarism, reneging on a commitment, writing a bad check, etc.

If the original Code were re-adopted, or some variation of it, the non-toleration clause would no longer be a part of the Honor Code. The non-toleration policy could continue, however, and perhaps more effectively. No longer would the stretching of the meaning of one's personal honor to include what one does about somebody else's violation be an integral part of the creed. Remember that 49% of the cadets indicated in an official survey that they wanted toleration removed as a personal violation. Those cadets may have been objecting to the specific point of stretched personal honor.

In lieu of the non-toleration clause, there is an alternative expression for the offense of tolerating another's honor violation. The word is, "misprision." To conceal a crime that one is not himself guilty of, to fail to report an offense committed by somebody else, to aid and abet a criminal in avoiding justice - that is misprision. A cadet who tolerates another's honor violation could be charged with misprision and, if convicted, levied heavy penalties.

As to the penalty for toleration, when the non-toleration clause was formally added to the Honor Code in 1970 in confirmation of informal practice of the non-toleration policy soon after the turn of the century, that fixed dismissal as the usual sentence. But the evidence suggests that toleration flourished despite the severity of the penalty. The result is as though a majority of the cadets challenged the justice of dismissal for every level of toleration. Dismissal may be appropriate for a tolerator who is an accessory before the fact of another's honor violation by cheating, and aids and abets the cheater's avoidance of justice after the fact. The tolerator who counsels a thief to return the stolen item, however, or who ignores his friend's unauthorized possession of a government pen, might be levied punishment short of banishment from the academy, in my view. Punishment to fit the crime of toleration at its various levels of seriousness might well increase cadet reporting of tolerators in a reversal of past indulgence of them.

Conclusion

The evidence suggests that the non-toleration clause of the Honor Code had not been effective in convincing more than a third of the cadets in 1974 to accept the non-toleration policy. Implementation of the non-toleration policy was ineffective for a period of ten years before 1976. Reliable and objective estimates of cadet skepticism about the non-toleration policy can be made available. Programs to increase cadets full acceptance and implementation of the non-toleration policy are proposed in order to reduce vulnerability of the Honor System to group cheating. If all actions fail to reduce vulnerability of the system to group cheating under the present Honor Code, an alternative criterion code would be useful to consider. The original code of the Military Academy is proposed for consideration if no other action produces relative invulnerability to groups of cadets violating the Honor System.

REFERENCES

Borman, F., Johnson, H. K., Pye, A. K., Tate, W. M., Walker, J. T., & Wilcox, H. S., Report to the Secretary of the Army, Assembly, March 1977.

Edney, J., Free Riders En Route to Disaster, Psychology Today, August 1979.

Gallup, G. Jr., Religion in America, 1979-80, Religion Research Center, 1979, 1-105.

United States Military Academy, History of the Honor Code and System, 1802-Present, Assembly, June 1977, 6-29.

United States Military Academy, The Honor Code and Honor System, USCC Pamphlet 632-1, June 1979.

For their criticisms of the ideas in this paper, I thank Martin Brown, John Carley, A. P. Clark, Ivan Farman, Andrew Goodpaster, Jack and Linda Heflin, Leif Hendrickson, Brad Hosmer, Charles Hosmer, Andrew Kinney, Dan McElheny, Robert McMahon, Thomas Moorman, N. A. Roth, Bill Sibert, Robert Stillman, and Clint True.

Improving Acceptance of Automated Counseling Procedures

James H. Johnson, Ph.D., Kathleen McGinley, Elizabeth Byrnes,

Steve W. Godin, Michael Bloomquist

Illinois Institute of Technology

Previous experience with innovative computerized systems suggests that the introduction of automated counseling procedures into the military environment will be attended with resistance and an initial low level of acceptance by users. However, user acceptance of these new systems can be improved if concepts from the technology of planned change and software psychology are integrated into the implementation process.

A consensual model of the change process is presented. This model structures organizational readiness, the change strategy, and acceptance as integrated variables to be considered in a successful installation.

A basic introduction to the principles of software psychology is presented. Past research in the installation of psychology related computer systems is considered in terms of these principles.

A structured approach for the installation of automated military counseling procedures is proposed. This approach takes into account issues of planned change and software psychology as well as previous practical experience in the area. A preliminary organizational assessment instrument for use in the installation of automated counseling procedures is described.

Improving Acceptance of Automated Counseling Procedures

James H. Johnson, Ph.D., Kathleen McGinley, Elizabeth Byrnes,

Steve W. Godin, Michael Bloomquist

Illinois Institute of Technology

In order to successfully install a computerized counseling system, it is necessary to consider a range of factors that can effect the project's outcome. The assessment of organizational readiness, structured implementation procedures, strategies for handling resistance, and system design features are all crucial variables in this process. The purpose of this paper is to beiefly review these areas to aquaint interested psychologists with a number of new developments in the field.

Organizational Readiness

The new technology of planned change emerged from a rational analysis of the forces that encourage and impede change. When an innovation is planned, the forces within the "target" organization must be understood and then developed to channel them for a successful implementation process. Davis (1971) used the acronym AVICTORY to describe the dimensions of organizational readiness that a "change agent" must consider in designing a plan for action. The target group must have the ability, that is the skills and resources, necessary to change. The innovation must be consistant with the values and operating style of the target. Information about the innovation must be provided. Other circumstances of the target should be considered so that the timing of the plan is optimal. Important personnel must feel the need or obligation to change. Sources of resistance to change must be anticipated. The yield, benefits and cost-effectiveness, should be clarified. Byrnes and Johnson (1981) constructed a questionaire based on Davis AVICTORY model to assess an organization's readiness for an automated system. Table 1 summarizes the dimensions of this questionnaire.

Table 1. A VICTORY Dimensions for Organizational Readiness for Automation

Category		
Ability	1	Emergence of a system's advocate, that is, someone who identifies the need for an automated system and will work to push it through.
	2	Power base of the organization is supportive of implementation.
	3	Availability of resources for automation.
Values	4	Administrative support and cooperation.
	5	Attitudes and beliefs of users toward automation.
	6	Organizational and administrative style.
	7	Organization's flexibility and openness to change.
Information	8	Availability of appropriate computer programs or computer science expertise.
	9	Presence of comparable manual system.
Circumstances	10	A climate of trust is needed so that communication can be facilitated.
Timing	11	High priority is being given to the automation project.
Obligation	12	Sources of motivation are present, internal and/or external, for automation.
Resistance	13	Fears of confidentiality, performance appraisal, and mechanization Expected negative consequences of automation.
Yield	14	General staff and administration both foresee the benefits of automation.

Change Strategies

Strategies to effect planned change should be chosen and combined according to the assessed characteristics of the target organization. Power strategies, that is the imposition of sanctions and rewards to change b ːavior, are recommended wehn staff commitment to the project low and/or the time frame is limited. Persuasive strategies, convincing the target group that the change serves its purposes, are suggested when the need or problem is not recognized or the innovation is not considered acceptable. Reeducational strategies, aimed at changing attitudes and values, are useful when commitment to the change is minimal and a long time frame is possible. Facilitative strategies which provide assistance for easier implementation, are recommended as an augmentation to other strategies (Zaltman & Duncan, 1977). Table 2 suggests types of issues to be considered in choosing strategies.

Previous reports about attempts to install automated procedures for psychological purposes suggest that failure to consider such factors has often led to serious installation problems and even rejection of innovative systems (Johnson, Williams, Giannetti, Klingler, & Nakashima, 1978; Hedlund, Sletten, Evenson, Altman & Cho, 1977; St. Clair, Siegel, Caruso & Spivack, 1976; Leader & Klein, 1977; Landau & Wilkes, 1975).

Software Psychology

To facilitate acceptance of an automated system the users' comfort and ease in operating the computer are important. Certain design features can be varied to develop a system appropriate to the users' level of experience with computers. More "powerful" commands and more "flexible" commands (that is commands that permit the user to perform an operation with fewer instructions or with a variety of types of commands, respectively) are less demanding of users (Ramsey & Atwood, 1979). Such command patterns should be balanced by a low level of "complexity" so that only a few commands need to be used at one time (Stewart, 1976). Computer-initiated dialogue reduces the amount of learning necessary to operate the computer, thereby reducing errors. Examples of computer-initiated dialogue are "yes or no" questions, "form-filling" (short answer), and "menu selection" (multiple choice) (Thompson, 1971; Martin, 1973). Although appropriate for inexperienced users, such procedures restrict the communication of information. To accomodate increasingly experienced users a "dual mode dialogue" which allows switching to a user-initiated approach is recommended (Pew & Rollins, 1975).

Programs can be designed to give messages to the user to guide interactions. Abmbiguous input can be resolved by presenting the user with alternative interpretations from which he/she can choose further commands (Plath, 1972; Codd, 1974). Patterns of errors can be detected and "instruction" for correcting such errors can be presented (Rouse, 1977). The pace of all such guidance should be programmed to be under the user's control (Goodwin, 1974).

Table 2. Sample Questions for Surveying Change Strateties Used During Implementation Processes

Power/Coercion

- Were top administrators used effectively before and during the implementation process?
- Was the administrative staff well educated about computer approaches?
- Would the chief administrator have penalized an employee who failed to cooperate with the needs of the automated system?
- Was the chief administrator propared for resistance to change prior to the implementation process?

Persuasion

- Were staff involved in all stages of the design and implementation process?
- Were there full-time liaison personnel to work with the staff?
- Were computer science trained personnel carefully exposed to psychologist's concepts prior to beginning work on the project?
- Were staff opinion leaders identified and involved in the installation of the computer system?

Reeducation

- Were numerous meetings and workshops held between the prospective users and project staff?
- Were key staff members given training in data processing?
- Were members sent to other facilities with automated systems for purposes of education?
- Was a computer project staff member assigned responsibility for user education?

Facilitation

- Was the project given adequate financial support?
- Did the design of the system attempt to minimize organizational change?
- Were chief administrators supporting the project?
- Were additional staff provided to allow sufficient total involvement during implementation?

When recruits or enlisted personnel are to use the computer, special care must be taken to ensure that collected data are accurate. Personnel unable to use a computer should be screened out. Johnson et al (1977) have constructed such a screening device. A period of orientation and practice should be allowed before data collection begins. Standardized responses and specially marked or color coded keys are suggested to make computer interaction easier. Design should allow for storage of partial data if a user is unable to complete a program in a single setting (Cole, Johnson & Williams, 1976).

A design that allows staff easy access to stored information expands the usefulness of the system for purposes of monitoring, documentation, program evaluation, and planning. System monitoring programs provide records of errors for ongoing evaluation and refinement of the system.

Approaches found to facilitate interpretation of output data are: graphics (Schniederman, 1980), data summary displays (Ramsey & Atwood, 1979), coding by color, brightness, size and grouping (Stewart, 1976; Teichner, Christ & Corso, 1977), and flashing important data (Smith & Goodwin, 1971). Output of printed reports can be varied in size, type, and style according to the users' needs. Alternative sentence stems can be varied so that each report is structurally unique. Retrofit programming techniques (fitting data from disparate collection procedures backwards around key data) produces unified reports that readers seem to like better (Johnson et al., 1977).

Important physical design features include: alternative keyboards to accomodate typers and non-typers (Schniederman, 1980; Hirsch, 1976), computer response time appropriate to the user's perception of task difficulty (Schniederman, 1980), yellow displays to reduce eye strain (Ramsey & Atwood, 1979), and elimination of flicker which causes fatigue and dissatisfaction (Dill & Gould, 1970). Detailed design guidelines applicable to psychological settings have been developed by Hanson (1971), Wasserman (1973), Pew & Rollins (1975), Gaines & Racey (1975), and Schniederman (1980).

All factors facilitating user ease of operation and comfort with the computer should be considered to enhance efficiency and encourage acceptance of the system. These considerations are an important part of the facilitative implementation strategy.

Suggested Implementation Procedures

Keeping in mind the organizational readiness, strategy, and design features discussed above, the following step-by-step implementation procedure is suggested.

Step 1: Assessment. Systems developers should review the change technology and software psychology literature in order to acquire the orientation and skills (information and ability) necessary to contribute to the design of an implementation plan. The organizational readiness instrument (described above) is suggested to determine potential problems and resistances. The relationship between permanent personnel and temporary implementation staff should be clarified to alleviate resistance due to fears of loss of power (Jones, 1969). Meetings would be scheduled to obtain design input from staff (values), to deliniate the problems of the current system (establish obligation) and the rationale for an automated solution (demonstration yield). These procedures should ensure optimum participation by opinion leaders (who can best evaluate circumstances and plan timing). Opinion leaders need to become "insiders" who are supportive of the proposed change and who can use their influence over other personnel to overcome resistances (Zaltman & Duncan, 1977).

Step 2: Plan design. The goals of the program must be clarified to make them consistant with the organization's values and to form the basis for later evaluations. Software and hardware should be selected based on organizational needs and the facilitation of user acceptance. Duties of all personnel must be deliniated. Procedures for feeding information about problems back to superiors should be described. In this manner a foundation can be developed for the rewards and sanctions of a power strategy.

Expressions of resistance are useful information in evaluating and refining design of the system. Resistance that is not openly expressed, but is only acted upon, is a serious implementation problem. Thus, it is necessary to define procedures for constructive expression of resistance. Planning and accepting resistance in a positive manner will increase involvement of personnel in the project, shift social support toward project success, discourage passive resistance, and provide necessary information for program evaluation and modification (Kelman & Warwick, 1973; Johnson et al., 1978).

Details about scheduling and training groups should be determined in advance to avoid disruption of organizational functioning and to form working units based on level of training needed and ability to provide support for the program's success.

Step 3: Training Personnel. Attendance at all training meetings should be required (power). Arrangements need to be made to adjust work loads during this period so that training dows not result in extra work (facilitation). Opinion leaders should be used in an orientation program to explain the rationale for the automated system (persuasive/reeducative strategy) and to allow staff to become familiar with the new equipment in a casual way (facilitation). Training sessions should be presented in small, successive steps to guarantee learing paced to the users' skill level. Extensive practice sessions often need to be scheduled (St. Clair et al., 1976).

Step 4: Ongoing training and evaluation. Temporary staff should be phased out and/or used on a consulting basis during this period. Monthly meetings and workshops need to be scheduled to solicit feedback regarding resistance, to inform personnel of progress and adjustments in the system, thus demonstrating yield and responsiveness to values expressed (Hedlund et al., 1977). These meetings provide a forum for continued persuasion and re-education of the staff over long periods of time, as well as to facilitate the program by continually increasing user skills.

Summary

To summarize, there are a number of factors related to the successful implementation of a computerized military counseling system that need to be considered. By assessing organizational readiness, and carefully structuring implementation procedures, a plan can be developed that will have the greatest probability of acceptance and success. It is recommended that all developers of computerized approaches pay careful attention to these factros if they wish maximum acceptance of their systems.

References

Byrnes, E. & Johnson, J.H. Change technology and the implementation of automation in mental health settings. Behavior Research Methods and Instrumentation, In Press, 1981.

Codd, E.F. Seven steps to rendezvous with the casual user, (Technical Report RJ-1333). San Jose, CA.: IBM Research Laboratory, January, 1974.

Cole, E.B., Johnson, J.H., & Williams, T.A. When psychiatric patients interact with computer terminal: Problems and solutions. Behavior Research Methods and Instrumentation, 1976, 8, 92-94.

Davis, H.R. A checklist for change. In National Institute of Mental Health. A Manual for Research Utilization, Washington, D.C.: Government Printing Office, 1971.

Gaines, B.R., & Facey, P.V. Some experience in interactive system development and application, Proceedings of the IEEE, 1975, 63, 894-911.

Goodwin, N.C. Intro: In which a smart terminal teaches its own use. (Technical Paper MTP-150). Bedford, MA: MHRE Corp., March, 1974.

Hanson, W.J. User engineering principles for interactive systems, Proceedings of the Fall Joint Computer Conference, 1971, 39, 523-532.

Hedlund, J.L., Sletten, I., Evenson, R., Altman, H. & Cho, D.W. Automated psychiatric information systems: A critical review of Missouri's Standard System of Psychiatry. Journal of Operational Psychiatry, 1974, 25, 103-104.

Hirsch, R.S. Human factors in man-computer interfaces, San Jose, CA. IBM Human Factors Center, 1976.

Johnson, J.H., Williams, T.A., Giannetti, R.A., Klingler, D.E. & Nakashima, S.R. Organizational preparedness for change: Staff acceptance of an on-line computer assessment system. Behavior Research Methods and Instrumentation, 1978, 10, 186-190.

Kelman, H.C. & Warwick, D.P. Bridging micro and macro approaches to social change: A social-psychological perspective. In G. Zaltman (Ed.), Processes and phenomena of social change. New York: John Wiley & Sons, 1973.

Landau, B. & Wilks, C. Development of an automated data system for the psychiatric service of the juvenile and family court. In B. Willer, G.H. Miller & L. Cantrell (Eds.), Information and feedback for evaluation, Toronto: The Communications Department, York University, 1975.

Leader, M.A. & Klein, D.F. Reflections on instituting a computerized psychosocial history in a clinical facility. Comprehensive Psychiatry, 1977, 5, 489-496.

Martin, J. Design of man-computer dialogues. Englewood Cliffs, New Jersey: Prentice-Hall, 1973.

Pew, R.W. & Rollins, A.M. Dialog specification procedures (Rev. Ed. (Technical Report No. 3127). Cambridge, MA: Bolt, Beranek & Newman, Inc., September, 1975.

Plath, W.J. Restricted English as a user language. Yorktown Heights N.Y.: IBM Watson Research Center, No., 1972.

Ramsey, H.R., & Atwood, M.E. Human factors in computer systems: A review of the literature. (Technical Report No. SAI-79-111-DEN) Englewood, Co.: Engineering Psychology Programs, Office of Naval Research, September 1979.

Rouse, W.B. Human-computer interaction in multitask situations. IEEE Transactions on Systems, Man and Cybernetics. 1977, 5, 38

Schniederman, B. Software psychology: Human factors in computer an information systems. Cambridge, MA.: Winthrop Inc., 1980.

Smith, S.L. & Goodwin, N.C. Blink coding for information display. Human Factors, 1971, 13, 283-290.

St. Clair, C., Siegel, J., Caruso, R. & Spivack, G. Computerizing a mental health center information system. Mental Health, 1976, 4, 10-18.

Stewart, T.F.M. Ergonomic aspects of man-computer problem solving. Applied Ergonomics, 1976, 7, 137-146.

Teichner, W.H., Christ, R.E., & Corso, G.M. Color research for visua displays. (Report No. ONR-CR213-102-4F). Las Cruces, NM: New Mexico State University, Department of Psychology, June 197

Thompson, D.A. Interface design for an interactive information retri system: A literature survey and a research system description. Journal of the American Society for Information Science, 1971, 22, 361-373.

Wasserman, A.I. The design of "idiot proof" interactive programs. AFIPS Conference Proceedings, 1973, 42, M34-M38.

Zaltman, G. & Duncan, R. Strategies for planned change. New York: John Wiley & Sons. 1977.

Enlistment Screening Test to Predict Army Aptitude Composite Scores

Richard A. Kass, Mary Weltin, Leonard Seeley and Hilda Wing

U.S. Army Research Institute
for the Behavioral and Social Sciences

All applicants for military service are required to achieve a minimum score on the Armed Forces Qualification Test (AFQT) to be eligible for enlistment. The AFQT is a composite score derived from four of the subtests of the Armed Services Vocational Aptitude Battery (ASVAB). Applicants for the Army also have to achieve minimum scores on Army Aptitude Area Composites which are various combinations of the ASVAB subtests.

The traditional Enlistment Screening Test (EST) used by recruiters for all military services to screen service applicants for potential failure has been designed to predict the AFQT score (Bayroff, Thomas & Kehr, 1959; Jensen & Valentine, 1976; Mathews, 1981; Morton & Houston, 1957). The U.S. Army Research Institute (ARI) has developed a Pre-enlistment Recruiting Test (PERT) to predict ASVAB Army Aptitude Area Composite scores, as well as scores on the AFQT. The PERT mini-battery is a shorter but parallel version of the operational ASVAB. PERT scores will be used within the Army's new Joint Optical Information Network (JOIN) system, a system using mini-computers with video display capabilities for use in Army recruiting stations. PERT-derived Aptitude Area scores will allow a recruiter to discuss particular Military Occupational Specialties (MOS) for which an applicant might be most qualified.

This paper presents preliminary results of the criterion-related validation of the PERT against the operational ASVAB.

Method

Instrument

The PERT was developed during the fall of 1980. Table 1 presents the correspondence between the subtests of the PERT and the operational ASVAB 8/9/1 The 80 items comprising the eight non-speeded PERT subtests were collected as follows. Forty items were from ASVAB Form 2 and four items from ASVAB Form 1. Twenty six items were original and ten Paragraph Comprehension items were obtai from a cross-section of three experimental Enlistment Screening Test (EST) booklets (Mathews, 1981). All of the 72 Coding Speed (CS) items were original. None of the PERT items is in the operational ASVAB.

There are three differences between the PERT and ASVAB. First, the PERT has no equivalent subtest for Numerical Operations (NO). The Coding Speed (CS) subtest in PERT substituted for both the CS and NO in the operational ASVAB. While the correlation between CS and NO in the operational ASVAB is only .64, the patterns of correlations between these two speeded subtests with the non-speeded ASVAB subtests are quite similar (see Table K-3 in Sims and Truss, 1980) Second, the PERT subtests have fewer items than the ASVAB subtests. Third, the items for the PERT subtests are not presented as separate content areas nor are they independently timed. The one exception is the PERT subtest for Coding Speed which is presented in a separate booklet (Test Book II) and has a five minute time limit. The other PERT subtest items are sequentially presented in Test Book I such that three items from a subtest are followed by three items from the next subtest. This sequencing continues until nine items from each of the subtests are presented. The last eight items of the 80 items in Test Book I are the successive presentations of the 10th items in each of the eight PERT subtests. The major advantage is that the recruiter does not need to monit time limits nor give specific directions for each subtest. The examinee is provided with initial directions and example problems and allowed to complete Test Book I (all subtests except CS) in 50 minutes. In the event a slower examinee does not finish, each subtest should be affected about equally. The examinee is then provided with Test Book II, and allowed five minutes to answer the 72 CS items.

Procedure

In March of 1981, the U.S. Army Recruiting Command (USAREC) tasked each of the 57 District Recruiting Commands (DRC's) to select 10 of the recruiting stations in their districts to participate in the research project. This would yield a total of 570 recruiting stations. They were selected in the following manner. First, all recruiting stations which had station commanders who were not on production were identified. (Non-production station commanders are not directly involved in soliciting applicants). Second, within each DRC, up to 10 recruiting stations were randomly selected. The DRC's and their designated recruiting stations participated in the study until each station commander tested 15 applicants and forwarded to the DRC 15 complete and usable answer sheets. Each DRC then forwarded all answer sheets to ARI.

Table 1

Correspondence Between the PERT and

ASVAB Subtests

Subtest Name	Content	Number of Items ASVAB	PERT
Word Knowledge (WK)	Understanding the meaning of words, i.e. vocabulary	35	10
Arithmetic Reasoning (AR)	Word problems emphasizing mathematical reasoning rather that mathematical knowledge	30	10
Paragraph Comphrehension (PC)	Understanding the meaning of paragraphs	15	10
Numerical Operations (NO)	A speeded test of four arithmetic operations, i.e., addition, subtraction, multiplication and division	50	None
Mechanical Comprehension (MC)	Knowledge of general mechnical and physical principles	25	10
Electronics Information (EI)	Knowledge of electronics and radio principles	20	10
Auto-Shop Information (AS)	Knowledge of auto mechanics, shop practices and tool functions.	25	10
Mathematics Knowledge (MK)	Knowledge of algebra, geometry and fractions	25	10
General Science (GS)	Knowledge of the physical and biological sciences	25	10
Coding Speed (CS)	A speeded test of matching words and numbers	84	72

During the last week in April, each of the designated station commanders was sent a packet of materials which included: a letter explaining the purpose of the research project; PERT Test booklets and IBM scorable answer sheets; test administration instructions. The station commanders were instructed to administer the PERT to all applicants. An individual was defined as an applicant when a recruiter completed a USAREC Form 714 for the individual. (A Form 714 is completed on applicants who are intending to be tested by the operational ASVAB). After that time, and before the applicant received the operational ASVAB, the station commander administered the PERT and had the applicant record all responses on a single answer sheet. The testing room normally used by the recruiters at the recruiting station was used for this purpose.

Subjects

This procedure yielded 2,921 answer sheets returned to ARI. Nineteen of these were dropped due to missing social security numbers. The distribution of the remaining 2,902 responses by DRC and Regional Recruiting Command (RRC) are presented in Tables 2 and 3, respectively. A large portion of the PERT testing was completed in May with the remainder finishing in June. Accordingly, for data analyses the May applicants were treated as the developmental sample and the June applicants as the validation sample. The criteria, the actual ASVAB test results, were obtained from the Military Processing Command (MEPCOM) for the May and June applicants. When the social security numbers were matched against those on the MEPCOM tapes, 1,058 and 478 matches were found for the May and June applicants respectively. The applicants who had results for tests other than ASVAB 8/9/10 were dropped, yielding developmental and validation samples of 1,047 and 473 respectively. The distribution of these two samples with respect to the DRC where PERT was administered is also presented in Tables 2 and 3.

An inspection of Table 2 indicates that the applicants tested were nationally distributed with no apparent regional biases. Table 3 summarizes these data with respect to the five recruiting regions. Comparing the percent of matched returns between the developmental and validation sample in each region highlights a problem with using month of testing as the basis for dividing the total sample into a developmental and validation sample. Since the majority of the applicants in the Northeast and Southeast Regions were tested early (i.e., in May) these two regions are over-represented in the developmental sample and under-represented in the validation sample. Just the opposite occurred for the Southwest, Midwest and Western Regions. While this unequal distribution is unfortunate, it will serve to make the validation of the regression weights computed in the developmental sample a more stringent test, since the validation sample may be less similar to the developmental sample. This will tend to decrease the size of the cross-validated Regression Coefficients. To ascertain the extent to which the developmental and validation samples differed, demographic characteristics and AFQT scores were examined for each sample.

Table 2

The Number of Applicants Tested on the PERT in Each District Recruiting Command (DRC)

District Recruiting Command	Number of Applicants: Tested with PERT	Matched with ASVAB: Developmental	Matched with ASVAB: Validation
Northeast Region			
Albany	28	13	--
Baltimore	33	15	4
Boston	33	9	1
Concord	25	16	--
Harrisburg	49	30	1
New Haven	28	7	--
Long Island	103	37	1
Newburg, NY	39	22	--
Ft. Monmouth, NJ	--	--	--
Niagara Falls	67	28	--
Philadelphia	81	28	2
Pittsburg	29	14	--
Syracuse	87	31	--
Total	602	250	9
Southwest Region			
lbuquerque	31	11	3
allas	67	29	6
enver	18	1	3
ouston	43	17	8
ackson	34	13	5
ansas City	46	28	10
ittle Rock	7	8	15
ew Orleans	31	18	6
klahoma City	64	18	17
an Antonio	43	15	9
Total	384	158	82

District Recruiting Command	Number of Applicants: Tested with PERT	Matched with ASVAB: Developmental	Matched with ASVAB: Validation
Southeast Region			
Atlanta	34	14	--
Beckley	--	--	--
Charlotte	75	24	3
Columbia	76	17	3
Jacksonville	53	29	7
Louisville	12	8	5
Miami	81	40	6
Montgomery	58	18	5
Nashville	90	27	13
Raleigh	108	37	2
Richmond	15	6	7
San Juan	85	9	19
Total	687	229	70

Table 2 (continued)

District Recruiting Command (DRC)	Number of Applicants			District Recruiting Command (DRC)	Number of Applica		
		Matched with ASVAB				Matched with AS	
	Tested with PERT	Develop- mental	Validation		Tested with PERT	Develop- mental	Va
Midwest Region				Western Region			
Chicago	27	6	9	San Francisco	72	32	
Cincinnati	57	23	15	Honolulu	--	--	
Cleveland	105	31	31	Los Angeles	40	14	
Columbus	99	44	22	Phoenix	72	17	
Des Moines	16	9	7	Portland	23	5	
Detroit	27	11	7	Sacramento	59	10	
Indianapolis	60	27	16	Salt Lake City	25	11	
Lansing	39	11	11	Santa Ana	89	21	
Milwaukee	72	30	17	Seattle	45	19	
Minneapolis	56	21	17	Total	425	129	
Omaha	36	17	12				
Peoria	78	24	26				
St. Louis	56	21	11				
Total	728	275	201				

Note: The number of applicant test results with unidentifiable DRC was 76 for the total number of applicants tested, 6 for the developmental sample, and 4 for validation sample.

Table 3

The Number of Applicants Tested on the PERT in Each Regional Recruiting Command (RRC)

Regional Recruiting Command (DRC)	Number of Applicants		
		Matched with ASVAB	
	Tested with PERT	Develop-mental	Validation
Northeast	602	250 (42%)	9 (1%)
Southwest	384	158 (41%)	82 (21%)
Southeast	687	229 (33%)	70 (10%)
Midwest	728	275 (38%)	201 (28%)
Western	425	129 (30%)	107 (25%)
	2,826	1,041 (37%)	469 (17%)

The distribution of both samples with respect to the demographic variables of Gender, Race and Education information available from the MEPCOM tapes, is presented in Table 4. The validation sample includes a slightly higher proportion of whites than the developmental sample, i.e., 71% vs. 61%. The proportion of males in both samples is identical, 78%. The distribution of applicants by education for the developmental sample is 41% for high school graduates, 14% for high school seniors and 45% for those with General Educational Diplomas (GED). The corresponding distribution fcr the validation sample is 47%, 11% and 42%. Thus, the developmental and validation samples are similar with respect to demographic characteristics.

A critical dimension of both samples is the range of scores on the AFQT portion of the operational ASVAB. Table 5 reveals that the developmental and validation samples are similar in respect to AFQT scores. Both samples include a large proportion (71% and 67%) of applicants who scored below the 50th percentile. These proportions are fortuitous for our purposes because the PERT predictions will be most useful for those applicants in the lower ability levels who may not qualify for all MOS. In general, comparisons between the developmental and validation samples indicate that there does not appear to be any major difference between the two samples, even though the two samples were not equally distributed among the five Recruiting Regions.

<u>Analyses</u>

To examine some psychometric properties of the PERT, subtest scale means, standard devitions and Cronbach's coefficient alpha's, an index of the internal consistency reliability, were computed in the developmental sample. The validation of the PERT was accomplished by computing eleven separate regression equations in the developmental sample. For each regression analysis, the PERT subtest raw scores served as the predictor variables and each of the ten Army Aptitude Area Composites and the AFQT successively served as the criterion variable. Two features of these regression analyses need to be explicated. First, the PERT was designed to predict Army Area Aptitude Composites, not ASVAB individual subtest scores. The ten Area Aptitude Compostles with corresponding MOS are presented in Table 6. Second, one PERT subtest, Coding Speed (CS), was excluded as a predictor in the regression analyses. Preliminary analyses of the distribution of CS scores showed many high scores. The distribution indicated that about half of the applicants were apparently allowed to respond beyond the 5-minute time of the subtest, thus invalidating the results.

Once the regression equations were computed in the developmental sample, the regression coefficients (non-standardized beta weights) were used to compute predicted ASVAB composite scores in the validation sample. The correlation between the predicted and actual ASVAB composite scores in the validation sample constituted the "validated" multiple R.

Table 4

The Distribution of the Developmental and Validation Samples with Respect to Gender, Race and Education

			Developmental	Validation
White	HS Graduate	Males	191	114
		Female	61	41
	HS Seniors	Male	56	28
		Female	21	9
	GED	Male	227	124
		Female	30	19
			636 (61%)	335 (71%)
Black	HS Graduate	Male	103	37
		Female	60	19
	HS Senior	Male	43	5
		Female	20	9
	GED	Male	108	43
		Female	27	5
			361 (34%)	118 (25%)
Other	HS Graduate	Male	15	8
		Female	4	2
	HS Senior	Male	5	1
		Female	1	0
	GED	Male	22	8
		Female	3	1
			50 (5%)	20 (4%)
		Total	1,047 (100%)	473 (100%)

Table 5

The Distribution of AFQT Scores
on the Operational ASVAB for the
Developmental and Validation Samples

AFQT Mental Category	AFQT Percentile Score	Developmental Sample (n = 1,047) Percent	Developmental Sample (n = 1,047) Cum. Percent	Validation Sample (n = 473) Percent	Validation Sample (n = 473) Cum. Percent
1	93-99	2	100	1	100
2	65-92	18	98	20	99
3A	50-64	9	80	12	79
3B	31-49	21	71	26	67
4A	21-30	14	50	13	41
4B	16-20	12	36	12	28
4C	10-15	15	24	9	16
5	1-9	9	9	7	7

Table 6

The Area Aptitude Composite Prerequisite for Major Groups of Army MOS

Area Aptitude Composite	Component ASVAB Subtests[1]	Military Occupational Specialties (MOS)
CO (Combat)	AR+AS+MC+CS	Infantry, Armor, Combat Engineer
FA (Field Artillery)	AR+MK+MC+CS	Field Cannon and Rocket Artillery
EL (Electronics Repair)	AR+MK+EL+GS	Missles Repair, Air Defense Repair, Electronics Repair, Fixed Plant Communications Repair
OF (Operators and Food)	NO+VE[2]+MC+AS	Missles Crewmen, Air Defense Crewmen, Driver, Food Services
SC (Surveillance and Communication)	NO+CS+VE+AS	Target Acquisition and Combat Surveillance, Communication Operations
MM (Mechanical Maintenance)	NO+EI+MC+AS	Mechanical and Aircraft Maintenance, Rails
GM (General Maintenance)	MK+EI+GS+AS	Construction and Utilities, Chemical, Marine Petroleum
CL (Clerical)	NO+CS+VE	Administrative, Finance, Supply
ST (Skilled Technical)	VE+MK+MC+GS	Medical, Military Policeman, Intelligence, Data Processing, Air Control, Topography and Printing, Information and Audio Visual
GT (General Technical)	VE+AR	Not currently used for classification into a particular MOS

Note: [1] standardized subtest scores are used in computation of composites.
[2] Verbal (VE) is a standard score conversion of the sum of the raw scores for WK and PC.

Results

Table 7 presents the means, standard deviations and reliability coefficients for the PERT subtests. Mathematics Knowledge (MK) was the most difficult of the PERT power subtests. Its mean of 3.5 is substantially lower than the means for th other subtests. The mean and standard deviation for Coding Speed (CS) is misleadi since, as mentioned previously, many applicants were apparently allowed to go beyond the specified time limit. The reliabilities of the power scales are quite adequate for ten-item scales.

Results of the regression analyses are presented in Table 8. The multiple R's in the developmental sample were quite high, ranging from a low of .73 for the Clerical (CL) Aptitude Area to a high of .82 for the Skilled Technician (ST) and General Technical (GT) Aptitude Areas and the AFQT. When the regression coefficients were validated in the second sample, multiple R's decreased only slightly, with the exception of predicting CL, which decreased from .73 to .63. In general, however, the PERT appears to be a useful predictor of Army ASVAB composites.

An examination of the standardized beta weights for the PERT subtests (Table 8) reveals that the most useful subtests were Arithmetic Reasoning (AR), Word Knowledge (WK), Mechanical Comprehension (MC) and Auto/Shop Information (AS) followed by Math Knowledge (MK), Paragraph Comprehension (PC) and General Science (GS). The PERT subtest Electrical Information (EI) did not achieve a single beta weight above .10. Comparisons of the distribution of beta weights across the PERT subtests in each regression equation with the actual corresponding ASVAB subtests used to compute ASVAB composites (indicated by underlinings in Table 8), suggests only a moderate relationship. This may be the result of sample fluctuations as the correlations of the operational ASVAB subtests are moderate to high (see Sims and Truss, 1980; Table K-3). It must also be noted that the PERT subtests are typically based on half (or less) of the number of items than the ASVAB subtests and, consequently, have lower reliabilities. The ASVAB subtest scale reliabilities range from .80 to .93 (Ree, Mullins, Mathews and Massey, in press). This lower reliability of the PERT scales would account for some of the coefficient attenuation.

Summary and Conclusion

In this investigation, a newly developed test was validated. The PERT was designed to be used by recruiters to test Army applicants in order to obtain an indication of the applicant's eligibility for enlistment as well as for specific MOS. One of the requirements for MOS entry is a qualifying score on the relevant Army Aptitude Area Composite. The PERT responses of 1,047 May 1981 Army applicants were used to develop regression weights to predict their ASVAB composite scores. The prediction equations were validated in an independent sample of 473 June 1981 Army applicants. The results from this research demonstrate that the PERT would be a useful tool for recruiters. The final step remaining before implementation of the PERT as an Army recruiting tool is to equate the PERT scales to the ASVAB scales. Recruiters need to know what ASVAB values are predicted by which PERT values. This step will be compeleted in the near future.

Table 7

Means, Standard Deviations and Reliabilities of PERT Subtest Scales

PERT Subtest [1]	Mean	Standard Deviation	Coefficient Alpha
Word Knowledge (WK)	5.6	2.6	.74
Arithmetic Reasoning (AR)	5.1	2.6	.76
Paragraph Comprehension (PC)	5.7	2.6	.73
Mechanical Comprehension (MC)	5.4	2.3	.66
General Science (GS)	5.2	2.2	.62
Electronics Information (EI)	5.0	2.3	.60
Mathematics Knowledge (MK)	3.5	2.2	.64
Auto/Shop Information (AS)	4.9	2.5	.69
Coding Speed (CS)[2]	36.0	18.1	

Note: [1]All PERT subtests have 10 items apiece, except CS which has 72.

[2]Coefficient Alpha was not computed for CS since it is a speeded test.

Table 8

Regression Weights and Multiple R's for Predicting Army Aptitude Area Composites from PERT Subtest Scores

Army Aptitude Area Composite	Multiple R Developmental Sample	Multiple R Validation Sample	WK	AR	PC	MC	EI	AS	MK	GS	CS
Combat (CO)	.78	.74		.31		.18		.24			—
Field Art. (FA)	.78	.72		.35		.17			.20		—
Electronics (EL)	.80	.77	.21	.28		.12	—	.12	.12		
Oper/Foods (OF)[3]	.80	.76	.18	.24	—	.16		.22			
Surv/Comm (SC)[3]	.77	.70	.17	.26	.15			.13	.11		
Motor Maint. (MM)[3]	.77	.75	.12	.22		.17	—	.28			
Gen. Maint. (GM)[3]	.78	.77	.21	.18	—	.13	—	.24	—	—	
Clerical (CL)[3]	.73	.63	.18	.26	.20				.18		—
Skilled Tech (ST)	.82	.77	.26	.20	—	.18		.11	.11	—	
Gen. Tech (GT)	.82	.75	.24	.35	.15	.10					
AFQT	.82	.76	.23	.35	.12				.16		

[1]ASVAB subtests used to compute composites are indicated by underlining.
[2]Standardized Beta Wts. less than .10 are omitted.
[3]ASVAB subtest NO also used to compute operational ASVAB Army composite.

References

Bayroff, A. G., Thomas, J. A. & Kehr, C. J. Evaluation of EST for Predicting AFQT Performance. (ARI Technical Research Report 1114). Alexandria, VA: Army Research Institute, February, 1959.

Jensen, H. E. & Valentine, Jr., L. D. Development of the Enlistment Screening Test - EST Forms 5 and 6. (AFHRL Technical Report 76-42). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, May, 1976.

Mathews, J. J. Development and Calibration of the Enlistment Screening Test (EST) Forms 81a and 81b. Paper presented at the meeting of the Military Testing Association, March, 1981.

Morton, M. A., Houston, T. J. & Bayroff, A. G. Development of Enlistment Screening Test, Forms 3 and 4. (ARI Technical Research Report 1102). Alexandria, VA: Army Research Institute, May, 1957.

Ree, M. J., Mullins, C. J., Mathews, J. J., & Massey, R. H. Armed Services Vocational Aptitude Battery: Item and factor analysis of Forms 8, 9, and 10. (AFHRL-TR-xx). Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division, in press.

Sims, W. H. & Truss, A. R. Normalization of the Armed Services Vocational Aptitude Battery ASVAB) Forms 8, 9, and 10 using a sample of service recruits. (CNA Research Contribution 438). Alexandria, VA: Center for Naval Analyses, December, 1980.

Footnotes

The assistance of CPT David Duke of the U.S. Army Recruiting Command in coordinating the distribution and collection of tests was invaluable. John Mathews with the Air Force Human Resources Lab was very helpful in disussion of data collection procedures and providing Paragraph Comprehension (PC) items. Dr. Kent Eaton, with ARI, provided critical institutional support and coordination. The opinions expressed are those of the authors and do not necessarily reflect those of the Army Research Institute or the Department of the Army.

Copies of this paper can be obtained from the senior author, U.S. Army Research Institue, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

The Great Training Robbery

Judah Katznelson, Research Psychologist
U.S. Army Research Institute for the Behavioral and Social Sciences
Alexandria, Virginia 22333

Abstract

Employee selection and training constitute the major methods available to an organization for improving the ability of the work force. In large part the major objective for both activities is identical, although they achieve this purpose quite differently. Selection seeks to enhance ability levels through a process of elimination. Selection procedures enable an organization to hire a greater proportion of high-ability employees than would otherwise be possible. Training, alternatively, seeks to achieve the objective by increasing the ability levels of the existing work force.

There are several reasons why most organizations will engage in both selection and training, even though the objective of the two is identical. For example, it is generally very difficult to achieve a highly skilled work force using just one of the procedures. All DoD elements use both procedures. However, by not going one step further and training employees to enable them to approach their jobs from a common organizational frame of reference, the organization has, in effect, robbed itself of valuable personnel and training resources. To prevent this, an innovative approach to training has enabled two DoD human factors organizations to increase their return for the training dollar and avoid the great training robbery.

No scene from prehistoric times is quite so vivid as that of the struggles of great beasts in the tar pits. In the mind's eye one sees dinosaurs, mammoths and sabretoothed tigers struggling against the grip of the tar. The fiercer the struggle, the more entangling the tar and no beast is so strong or so skillful, but that he ultimately sinks.

Training over the past years has been such a tar pit and many great and powerful beasts have thrashed violently in it. Many training approaches have emerged with working programs but few have met fully the goals and expectations. Large and small, massive and puny, approach after approach has become entangled in the tar. No one thing seems to cause the difficulty - any particular paw can be pulled away. But the accumulation of simultaneous and interacting factors brings slower and slower motion. Everyone seems to be surprised by the stickiness of the problem and it is hard to discern the nature of it. But we must try to understand the nature of it if we are to solve it.

This paper has three objectives. First, we will describe some basic dimensions of ability, focusing on differences in ability both within and between people. Second will be a description of how an organization can attempt to match the abilities of its employees with the ability requirements of its jobs. Finally attention is given to how an organization can attempt to upgrade the ability levels of its work force through its selection and training procedures.

We are all aware that group differences in physiological and psychological factors exist. What is often not so well known is that differences between individuals are generally much greater than differences between groups. Thus, knowing only one's group affiliation often tells us little about the individual even though we know the average performance of the group.

There are two types of individual differences that we must consider, interindividual and intraindividual differences. Interindividual differences pertain to differences between people - for example, differences in weight, intelligence and vision. On virtually every physical and psychological dimension, people demonstrate great variability.

Interestingly, however, the patterns of variability found among individuals on different characteristics tends to be quite similar. Specifically, the majority of persons tend to be arranged close to average on the characteristic, while relatively few people tend to be extremely high or low. If we were to take measurements on some characteristic from a fairly large number of persons, the frequency distribution would often appear as the classical bell-shaped curve.

These general observations should suggest to us something of the type of finding to be expected if we were to analyze the performance of a group of employees. If the distribution on some measure of performance varies markedly from the bell-shaped curve, caution is in order. Suppose, for example, that almost all of the employees are rated as high performers. In that case we might suspect that (1) the organization has been enormously successful in eliminating, through selection and/or training, the individual differences to be expected, or (2) bias has crept into the measurement system. The latter is frequently probable.

In contrast, intraindividual differences occur within individuals and have to do with the relationship between two or more characteristics. For example, if we know that an individual is high in verbal ability, can we also assume that he is high in numerical reasoning ability? In general, the answer to questions of this sort is no. Although positive relationships do exist between certain characteristics within individuals, these relationships tend to be low.

Again, the implications for measuring performance are fairly obvious. We would expect little relationship between various performance subcomponents of an individual being observed. Someone who is a conscientious performer, for example, may have only average human relations skills and be quite unknowledgeable about the task. A high degree of correspondence between measured subcomponents taken on the same individual should make us suspicious of the process we are employing to get those measures.

In an organizational context, it is of little value to view the abilities we have been discussing in some absolute sense. We cannot say, for example, that an individual possessing third grade language capabilities is adequate or inadequate without additional knowledge about the job he is to perform. If the job requires only an ability to write one's name and read very simple instructions, the individual's language capabilities may be adequate. Indeed, evidence suggests that individuals with capabilities exceeding the requirements of the job may perform as inadequately as those who do not possess the requisite skills.

We think in terms of matching a person to the job he is to perform. Such a matching process, in turn, requires that we be able to measure the abilities of the individual and the ability requirements of the job. Individual abilities to perform on the job are often measured by performance appraisals designed to identify various traits or characteristics of workers. However, appraisal systems having worker's traits as their major focus have generally not been very satisfactory. This results partly from the difficulty of assessing individual traits by a procedure which requires one person to observe another. It is also partly a system fault - that is, the traits measured in many appraisal systems bear little obvious relation to successful performance of the task.

An alternative appraisal procedure which has gained increasing acceptance focuses on the behaviors rather than on the traits of an employee. Through the use of critical incidents and related techniques, efforts are made to identify behaviors which are closely related to either successful or unsuccessful task performance. Although more will be said of these procedures later, we wish to point out here that behaviors appear easier to observe than ability traits per se. Thus, successful performance appraisal systems are likely to measure behaviors which are one step removed from the direct measurement of abilities.

In the preceeding paragraphs ability was defined, the nature of individual differences in ability were identified, and the need to match individuals to tasks in terms of the abilities required was discussed. Now the focus turns to methods an organization has at its command to manipulate individual ability so that a congruent person-job match may be obtained. Our discussion will deal with two general procedures that exist, employee selection and employee training, or development.

A major way in which organizations attempt to manipulate the ability levels of their work force is through employee selection. The core problem of selection involves the identification of appropriate ability levels among job applicants. Appropriate ability is normally defined in terms of the types of skills required for successful performance of some task.

Thus, selection can be thought of as being concerned with the identification (prediction) of successful task performers prior to their employment and can be achieved by the use of one or more predictors (e.g., tests, interviews) to assess the probable future organizational success of job applicants.

Suppose, for example, that a group of sales aspirants applying to an organization are given a short general aptitude test and an interest inventory which measures preferences for various types of careers. After a sufficient time had elapsed, their ability to perform the job would be examined. The examination might take the form of a supervisory performance evaluation of each employee. This evaluation and the results from the two tests would then be compared. Suppose that this comparison showed that: (1) individuals preferring business and related careers on the interest inventory generally received higher evaluations than persons preferring nonbusiness-related careers, and (2) there was no relationship between aptitude test scores and supervisory evaluations.

The results obtained above could subsequently be used to assist in the selection of new employees. Specifically, future job applicants should be given the interest inventory, but not the aptitude test, since the latter showed no relation to job success. Efforts should be made to hire those applicants who indicate business and related career preferences on the interest inventory. This hiring procedure should result in the employment of a larger proportion of individuals who will contribute to the goals of an organization than would otherwise be the case.

The accuracy of the statement above depends on three important assumptions that require brief elaboration. First, for any selection procedure to be useful there must be more applicants than there are jobs. This is necessary so that the organization can choose among the applicants, accepting the "best" and rejecting the "poorest." In the illustration, this would consist of accepting individuals showing business and related career preferences on the interest inventory.

Second, we must assume that the conditions prevailing when the predictors were initially validated apply when they are used to make selection decisions. If they do not, the predictor-performance relationships observed in the validation study may inadequately describe the relationships under the changed conditions. We must assume, for example, that job applicants as a group remain essentially the same over time. This is likely to be a tenuous assumption if the job market fluctuates markedly. Additionally, we must assume that the content of the jobs involved remains stable over time. This, too, is a tenuous assumption in a technologically innovative organization.

The only certain test for the second assumption is a continual revalidating of our predictor instruments through time. We may expect to find some shifting about in the contribution of various predictors to successful selection. It should also be obvious from this discussion that the utilization of a predictor in our organization simply because it has demonstrated validity in some other organization is unwarranted. We cannot safely assume that the predictor-performance relationship will generalize to our situation no matter how similar the tasks and job applicants appear.

The final assumption is in some respects the most important. We must assume that the measure of job performance, in the illustration - a supervisory evaluation, "gets at" what we regard as important for the success of our organization. For example, do differences in evaluations reflect real performance differences, or do they instead merely reflect supervisors' preferences for various employees? For the purposes of this paper, let it suffice to say that employee selection constitutes one important personnel activity where adequate performance appraisal is crucial.

Training, or development, of employees is the second major method an organization employs to manipulate the ability levels of its work force. Training frequently involves new employees, but may also include existing workers whose skills are deemed insufficient for their current job or for a job to which they are promoted.

Like selection, training can be viewed as a process for manipulating skill levels. As such, training may be thought of as involving:

1. Identification of the skills to be learned through training;
2. Identification of participants to receive the training;
3. Development or selection of procedures which enable participants to learn efficiently the required skills;
4. Appraisal of the training procedures' effectiveness.

Once identifying what skills are to be learned, we can turn to an identification of the persons who would benefit from learning them. With new employees, everyone in the group may reasonably be included. This is particularly appropriate when the skills to be learned are relatively unique to the organization under consideration.

Bass And Vaughan (1966) suggest that any training technique be judged by how well it conforms to the findings from learning theory. As such, they suggest that an appropriate training procedure (pg. 86):

1. Provide for the learner's active participation;
2. Provide the trainee with knowledge of results about his attempts to improve;
3. Promote by means of good organization a meaningful integration of learning experiences that the trainee can transfer from training to the job;
4. Provide some means for the trainee to be reinforced for appropriate behavior;
5. Provide for practice and repetition when needed;
6. Motivate the trainee to improve his own performance;
7. Assist the trainee in his willingness to change.

An example of how this procedure was utilized can be seen from the efforts of two DoD human factors organizations: the US Army Human Engineering Laboratory, and the US Navy's Human Engineering Branch of the Pacific Missile Test Center.

These organizations faced a common personnel problem. Both groups hired people with various educational backgrounds. Not only did the educational accomplishments of the employees range from college to the post-graduate level, but the expertise ranged across various fields such as psychology, anthropometry, engineering and computer science. Like most organizations, these DoD components engaged in both selection and training to meet their personnel needs.

Because the Army and Navy are increasingly pleased with the success of self-paced, interactive instruction and programmed instruction texts, they collaborated on a computer-based training course that combined modern training procedures with an approach that allows the accomplishment of the seven steps mentioned above. Additionally, by recognizing and addressing the need to provide their people with a common technical frame of reference among human factors specialists - wherever they may be employed - these organizations have increased their return for the training dollar. By using this additional step, they have avoided the great training robbery that so often inhibits the potential for the maximum payoff of valuable personnel and training resources.

In conclusion, the most suitable evaluation from an organization's point of view will be direct evidence about employee performance. The basic question is, has the performance of the participants benefited from the training program. Thus in training, as in selection, the ultimate value can be determined only after employees' contributions to organizational objectives can be properly assessed and success can be realized when innovative approaches are used to increase the efficiency of good training and selection procedures.

The Use Of The Armed Services Vocational Aptitude Battery (ASVAB) As A Predictor Of Early Discharge From The U.S. Army

Edward Kemery, M.S.
Robert Pleban, Ph.D.
Beverly Williams, M.S.
Frederick Dyer, Ph.D.

U.S. Army Research Institute Field Unit
Ft. Benning, Georgia 31905

Abstract

Recent research by Maier and Grafton (1981) has demonstrated the effectiveness of aptitude composites from the ASVAB 8, 9, and 10 in predicting training success and job proficiency. The present study was concerned with assessing the ability of the ASVAB composite scores to discriminate individuals in Infantry training recommended for early separation under the Trainee Discharge Program (TDP) from successful trainees. The composite scores of 87 TDPs and 87 non-TDPs were compared. A discriminant analysis revealed that approximately 69% of the TDPs and 69% of the successful trainees could be correctly classified using a weighted combination of three ASVAB composites. The classification functions derived from the initial sample were applied to a second sample (N=80) and showed that approximately 70% of both the TDPs and successful recruits could be correctly classified. When these classification functions were applied to a more realistic sample (N=419) and success-failure base rate was taken into account, it was found that 97% of the successful recruits could be correctly classified, but only 3% of the TDPs. Reasons for the ASVAB's inability to accurately identify TDPs and directions for future research are discussed.

Introduction

Since the inception of the all volunteer Army in 1973, there has been an increasing focus on personnel attrition. Recent reports (e.g. Frank & Erwin, 1978, Youngblood, Laughlin, Mobley, & Meglino, 1980) indicate that approximately ten percent of all incoming enlistees do not make it through Basic Combat Training (BCT) and, One Station Unit Training (OSUT). Given the size of the total incoming enlistee population, such attrition rates represent a sizable cost to the Army in terms of both money and wasted training time.

Currently, recruits who are having problems (non-medical in nature) in OSUT and who have not responded to various motivational strategies (e.g. counseling) are recommended for early discharge under the Trainee Discharge Program (TDP). The Trainee Discharge Program was designed by the Department of the Army to permit the rapid separation of those individuals deemed unfit for military service. The existence of such a program is highly desirable from the standpoint of the unit's morale as well as one of training efficiency (i.e. the sooner undesirables are identified and removed, the less expense there is to the Army).

The rationale behind the Trainee Discharge Program is sound but can be quite costly. For example, during FY 1980 at Ft. Benning there were approximately 1400 TDPs; most of these individuals were identified by about the sixth week of training. When one considers the cost of putting one enlistee through OSUT at Ft. Benning during FY 1980 was $23,599.00, it becomes clear that a large amount of money is being lost by the Army on these unsuccessful recruits.

If, however, a significant number of these TDPs could be reliably identified and screened out before they begin OSUT, a substantial amount of time and money could be saved.

Several strategies have been suggested for identifying early attrites e.g. autobiographical data (Frank & Erwin, 1978); assessment of recruits expected satisfaction, intentions toward completing their enlistment, and their attraction to the role of being a soldier (Youngblood, Laughlin, Mobley, & Meglino, 1980). Early reports on these efforts are quite promising but they are still in the developmental stage and/or have yet to be implemented on a large scale.

Another potential selection device is the Armed Services Vocational Aptitude Battery (ASVAB). Although not designed specifically for this purpose, it does have the advantage of already being used on a large scale by the Department of Defense, and if found to be successful at identifying early attrites would represent a relatively cost free selection instrument.

The ASVAB was designed by the Department of Defense to assess mental qualification for enlistment and to classify accession into skill training programs. The ASVAB provides an Armed Forces Qualification score, which is already used to screen out applicants unqualified (ability wise) for enlistment. The most recent ASVAB, form(8/9/10), contains ten subtests and ten composites (i.e. Combat, Field Artillery, Electronics Repair, Operators and Food, Surveillance and Communications, Mechanical Maintenance, General Maintenance, Clerical, Skilled Technical, and General Technical). Each composite with the exception of General Technical is composed of from three to five subtests which were found to be the most valid predictors of success in the job training programs. The aptitude composites are used to determine the eligibility for assignment to job specialties. (See Maier & Grafton, (1981) for a more complete description of the subtests and the composites comprising the ASVAB).

The primary focus of this investigation however, was to assess the utility of ASVAB as a predictor of early attrition from the U.S. Army.

Method

Subjects

Scores on the ASVAB Composites, Ethnic Status, Performance-Oriented Infantry Qualification Test (POIQT) results, and graduation status (whether or not a recruit was a TDP) were obtained on 1094 men who were members of OSUT companies which began training between 1 September and 31 December 1980. Out of these 1094 recruits 170 were identified as TDPs.

Results and Discussion

During the time period sampled from, two different versions of the ASVAB were given (7b & 8, 9, 10). Fruchter and Ree (1977) have suggested that these are equivalent forms, and will be considered as such from a statistical perspective. A stepwise discriminant analysis (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975) based on 87 TDPs and 87 successful trainees (determined from passing POIQT scores) was performed using nine[1] ASVAB components as the discriminating variables. While the successful recruits scored significantly higher on all of these components, the results of the stepwise discriminant analysis indicated that the two groups could be optimally discriminated using three components: combat arms, clerical knowledge, and electronics information.

The chi-square associated with the resulting lambda of .8174 was highly significant, $X^2(3) = 34.368$, $p < .00001$, indicating that the three components could differentiate between the groups. It should be noted, however, that the large lambda signifies that the difference between the group centroids (.469 for the successful trainees and -.469 for the TDPs) was rather small.

The single discriminant function that was derived was also significant (by virtue of the significant chi-square) and took the following form:

Discriminant Score = -12.1111 + .03387(Combat) + .06076(Electronics) - .2705(Clerical)

Since the major interest was in the ASVAB's ability to identify TDPs, a separate "classification function" for each group was derived based on the three aforementioned discriminating variables (composites). The functions for each group were as follows:

<u>Successful Trainees</u>

Score = -80.660 + .56093(Combat) + .63416(Electronics) - .2705(Clerical)

<u>TDPs</u>

Score = -69.279 + .52909(Combat) + .57704(Electronics) + .25076(Clerical)

Each function was applied to both groups of enlistees, producing two classification scores per person. A case was classified into the group whose function yielded the highest score, since the scores have the property that the case resembles most closely that group on which it has the highest score (Klecka, 1980).

When these functions were applied to the derivation sample, 69.54% of all the recruits were correctly classified. Cell breakdowns showed that approximately 69% of the TDPs and 70.1% of the successful trainees were correctly identified.

[1] The General Technical (GT) score was omitted from the analysis due to its redundancy with several of the remaining components.

The same functions were then applied to a holdout (cross-validation) sample of TDPs (N=40) and successful recruits (N=40) and resulted in no loss of accuracy. In fact, 71.25% of all recruits in this sample were correctly classified. Cell breakdowns showed that 70% of the TDPs and 72.5% of the successful trainees were correctly identified.

In order to assess the effectiveness of these functions in a more realistic situation, a sample was created in which ten percent of the recruits were TDPs (N=42) and the remaining ninety percent were successful trainees (N=377). The classification functions were then applied to this second cross-validation sample; this time however, the functions were adjusted to take into account the prior probabilities of belonging to the successful trainee group (.90) or the TDP group (.10). The results of this validation effort indicated that 88% of all recruits were correctly classified. However, the individual cell breakdowns showed that 97% of the successful recruits could be correctly classified but only 3% of the TDPs (See Table 1). These results strongly suggest that the predictive utility of the ASVAB with respect to identifying early attrites is negligible.

Inspection of the distribution of discriminant scores provides some insight into the ASVAB's inability to identify early attrites. Figure 1 shows the distribution of discriminant scores for both the TDPs and successful trainees. It can be seen that the distributions of both groups overlap with each other to a considerable extent. The overlapping distributions and a 90% base rate of success in OSUT may explain in large part the inability of the ASVAB to discriminate the TDPs from the successful trainees.

Further analysis of the two distributions raises some other noteworthy points. Assuming that these sample distributions are representative of the actual population distributions for TDPs and successful trainees, one thing that stands out is that the population of TDPs appears more homogeneous and tends to cluster at the lower end of the continuum. There are, however, some notable exceptions. Approximately 22% of the TDPs fall in the mid to upper end of the continuum of discriminate scores. Also, of interest is that some of the successful recruits had scores at the extreme lower end of the continuum.

The cluster of TDPs at the lower end suggests that these individuals may be deficient in some core ability like reading comprehension and this alone could account for numerous problems in OSUT.

Nevertheless, it is also clear from Figure 1 that some people with these "deficiencies" do, in fact, make it through training. Several possibilities may account for this: the unreliability of the instrument, high levels of motivation that would overcome any learning deficit (i.e. halo effect), or certain dispositional characteristics (e.g. compliance) that would allow someone to go through training with a minimal amount of conflict. The possibility also exists that incompetent people are not being terminated from the Army during OSUT due to differential criteria for evaluating competence at the company level.

TABLE I

COMPARISON OF ACTUAL GROUP MEMBERSHIP WITH PREDICTED GROUP MEMBERSHIP FOR TDPs AND SUCCESSFUL TRAINEES*

		PREDICTED GROUP MEMBERSHIP	
ACTUAL GROUP	CASES	SUCCESSFUL	TDP
SUCCESSFUL	377	368 (97.6%)	9 (2.4%)
TDP	42	41 (97.6%)	1 (2.4%)

PERCENTAGE OF RECRUITS CORRECTLY CLASSIFIED: 88%

*PRIOR PROBABILITIES SET AT .90 (SUCCESS) AND .10 (FAIL)

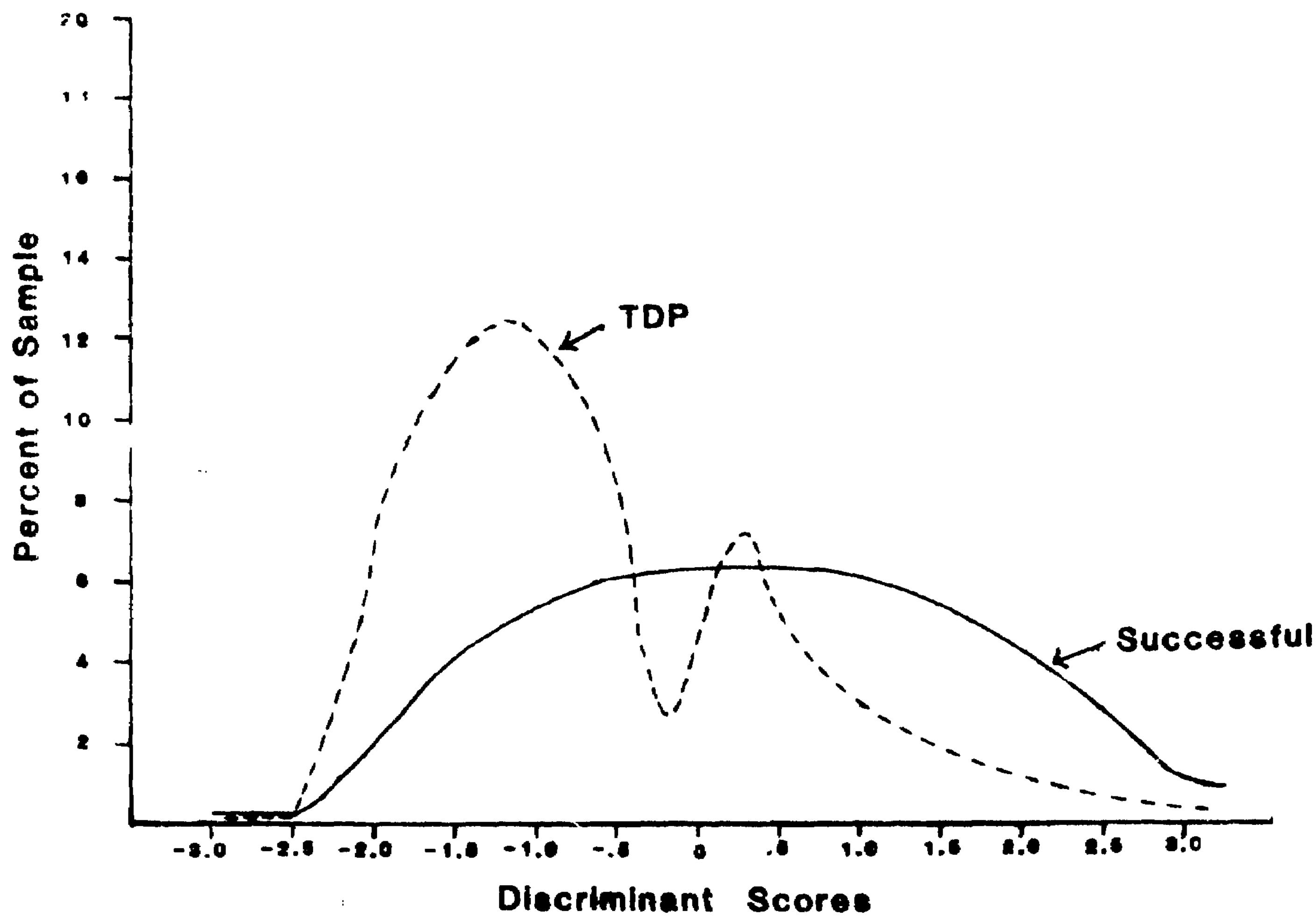


Figure 1. Distribution of scores for TDPs and Successful recruits.

The presence of TDPs at the mid to upper end of the distribution would seem to indicate though that a single core ability like reading comprehension (while important) is not the only reason why someone would be unsuccessful. It should come as no surprise that people might be unsuccessful for any number of reasons, some which are unrelated. There is no reason to expect, for example, that problems with authority figures should be related to reading skills.

Conclusion

The results of this study suggest that while the ASVAB may be a reliable predictor of job proficiency (Maier & Grafton, 1981) it does not appear to be a very effective predictor of attrition in basic training. It should be reiterated again, though, that the ASVAB was not designed with this purpose specifically in mind.

Nevertheless, this research effort does suggest a course for future research in this area. Although some aptitude/ability may be necessary to get through basic training, other factors which may be even more critical in determining trainee success are the motivation, interpersonal style and background of the recruits.

Thus, future efforts at identifying high risk recruits might be more profitably directed toward the following strategies (or combinations): intention and attitudinal measures (e.g. Youngblood, et al, 1980); autobiographical data (e.g. Frank & Erwin, 1978); and realistic job previews (e.g. Wanous, 1978).

Given the expenditures associated with those recruits who are separated from the Army prior to the completion of basic training, strategies for identifying those individuals before they enter basic training should be vigorously pursued and investigated.

References

Frank, B. & Erwin, F. The prediction of early Army attrition through the use of autobiographical information questionnaires. ARI Technical Report TR-78-A11, July, 1978.

Fruchter, D. & Ree, M. Development of the Armed Services Vocational Aptitude Battery: forms 8, 9, and 10. Education Development Corp. Technical Report 77-19, March, 1977.

Klecka, W. Discriminant Analysis. Beverly Hills, Ca.: Sage Publications, Inc., 1980.

Maier, M. & Grafton, F. Aptitude composites for ASVAB 8, 9, and 10. ARI Research Report 1308, May, 1981.

Nie, N., Hull, C., Steinbrenner, K., & Bent, D. Statistical Package for the Social Sciences. New York: McGraw-Hill, Inc., 1975.

Youngblood, S., Laughlin, J., Mobley, W., & Meglino, B. A longitudinal analysis of military recruit attrition: the first 25 months. Office of Naval Research Technical Report 11, February, 1980.

Wanous, P. Realistic job previews: can a procedure to reduce turnover also influence the relationship between abilities and performance? Personnel Psychology, 1978, 31, 249-258.

Kenaston, Alicia, US Office of Personnel Management, Washington, DC.
(Thurs. A.M.)

Applying Linguistic Theory to Job-Related Verbal Tests

Ongoing research at OPM has indicated the applicability of linguistic theory to the development of tests containing a verbal component. The ultimate aim of this research is a practical set of construction roles that will allow item writers to control the reading level of a test so that it (1) is job-related and (2) does not impede, to any significant level, performance on item types which are intended to test a non-verbal ability.

This paper describes a technique for measuring and controlling verbal difficulty based on applied linguistic theory. The technique has been used successfully to control the effect of verbal complexity on performance on a test of deductive reasoning. It is currently being evaluated to determine its usefulness in identifying the reading level of a job and in developing a job-related reading test.

The technique is explained, with examples, and its application to job-related verbal tests is discussed.

Applying Linguistic Theory to Job-related Verbal Tests

Alicia Kenaston
Personnel Research and Development Center
United States Office of Personnel Management, Washington, D.C.

Ongoing research at the Personnel Research and Development Center of the Office of Personnel Management focusses on developing a set of rules based on applied linguistic theory for the construction of verbal items. The ultimate aim of this research is a practical set of procedures that will allow item writers to control the verbal difficulty and make-up of an item in such a way that the verbal component of a test is both appropriate and job-related. The verbal component is appropriate if it does not impede, to any significant level, performance on item types which are intended to test a non-verbal ability, e.g., arithmetic reasoning. A job-related verbal component conforms to the level of linguistic ability required for successful performance on the job for which the applicant is being tested.

The initial thrust of this research has been to develop a method to control and quantify syntactic complexity at the sentence level. The research departs from the premise that the sentence is "the ordinary vehicle of linguistic reference" (Whorf, quoted in Bever, 1972). This implies that the receiver of a message, whether written or spoken, does not only decode the various lexical units with which he or she is confronted but must also recognize and decode the syntactic and semantic relationships which intertwine the lexical units in order to understand the message. Studies of child language acquisitio (for example, Bloom, 1970) have suggested that even the one-word utterances of children just learning to speak are actually sentences, loaded with syntactic and semantic information that the child has not yet learned to encode linguistically. Gesture and situational cues help to decipher these utterances. So, for example the utterance "mommy" has an almost unlimited number of potential meanings: "I see mommy;" "Mommy come here;" "I want mommy;" "This thing belongs to mommy;" etc.

The psycholinguistic theory upon which the method for measuring language complexity described here is based proceeds from the premise that the sentence is the basic unit for encoding and decoding a message. However, since the capacity of the short term memory is limited, a "pre-processor" of linguistic information has been hypothesized. That is, a sentence which is too long to be decoded in its entirety must be broken down into units which are individually deciphered and stored. The investigations of Bever (1972) enumerate three principal features of the pre-processing mechanism: (1) The clause is the primary perceptual unit; (2) Within each clause, semantic relations between major phrases are assigned; and (3) The clause is recoded into relatively abstract form to "make room" for the next clause. Bever's findings suggest, in part, that the number of clauses per sentence is a factor of sentence difficulty.

The Clause Analysis Technique

Cook (1975) has developed a clause analysis technique for the measurement of style complexity. This method has the advantages of being fairly easy to use while yielding detailed syntactic information. It can easily be employed by anyone with a strong grasp of traditional grammar and a mind open to a few insights from linguistic theory.

Independent of Bever's work, Cook designated the clause the basic unit of analysis. The technique involves three levels of clause structure: (1) the basic unit, (2) the information block, which consists of a main clause and the clause group clustered around it, and (3) the sentence, which may consist of one, or, in the case of compound sentences, more information blocks.

There are two major steps in the clause analysis technique, the first to identify the clauses and the second to compute the style indices, which are based on the three levels of clause structure described above. A clause contains a verb phrase, usually only one, although conjoined verb phrases are possible. The first rule for identifying clauses is thus to identify the verb phrases. Due to the application of transformational rules to the deep structure of a sentence, verbs are sometimes deleted from the surface representation of a clause. These verbless clauses include comparisons, postposed locative and adjective phrases, appositives, and so on.

Identifying Clauses. As stated earlier, the identification of clauses is essentially the identification of verb phrases. The following sentences can easily be analyzed into their clause components:

(1) They attended the conference that was sponsored by MTA.

(2) The investigator intended to analyze linguistic complexity.

(3) They arrived for an 8:00 am session drinking coffee.

(4) If you don't understand this, you can ask questions.

Each of the examples above contains two clauses. In the analysis of clauses, care must be taken not to separate auxiliary verbs from the lexical verbs to which they belong. Thus, in sentence 1 "was", the passive marker, is included as part of the verb phrase "was sponsored;" in sentence 4 "don't" and "can" are included as part of the verb phrases "don't understand" and "can ask". The auxiliary verbs are: (a) auxiliary "be" in the progressive tenses and the passive, (b) modal verbs such as "will," "can," "must," "should," etc., (c) the perfect marker "have", and (d) auxiliary "do." A verb phrase may have more than one auxiliary: "I didn't see the manuscript, but it might have been being typed."

Verbless clauses are only a little more difficult to uncover. The following examples do not exhaust the possible types of verbless phrases:

(a) comparative
(5) This year's conference is bigger than last year's (was).
(6) This workshop lasts as long as that one (lasts).

(b) postposed adjective and locative phrases
(7) They attended the conference (that was) sponsored by MTA.
(8) The papers (which are) under the heading 'Test Development' look interesting.

(c) appositives
(9) The headquarters of MTA, (which is) the sponsor of this conference, are located in Virginia.

Computing the Style Indices. Following the conventions of traditional grammar, one of the verb phrases is designated the main clause. In sentence 1, the main clause is "They attended the conference;" in sentence 4 the main clause is "you can ask questions." Generally, the main clause contains the subject of the sentence and the verb phrase closest to it.

After identifying the clauses, the analyst rewrites them, a clause to a line, indicating the main clause with the letter A, clauses subordinate to the main clause with the letter B, clauses subordinate to a B clause with the letter C, etc. This process is called reduction. The sample sentences may be reduced as follows:

(1) #A They attended the conference
B that was sponsored by MTA

(2) #A The investigator intended
B to analyze linguistic complexity

(3) #A They arrived for an 8:00 am session
B drinking coffee

(4) #B If you don't understand this
A you can ask questions

The symbol "#" is used to indicate the beginning of a sentence.

Compound sentences may also be reduced:

(10) #A The investigator presented a paper
+A and the participants asked questions

The symbol "+" indicates the boundary of an information block.

When the clauses have been identified and reduced, main and subordinate clauses indicated by the appropriate letter, the analyst may begin the second step, the computation of language complexity. Cook describes three indices of complexity: (1) average sentence length (ASL) -- the average number of clauses per sentence, (2) average block length (ABL) -- the average number of clauses per information block, and (3) average clause depth (ACD) -- the average depth of embedded clauses. The "depth" of an embedded clause is determined by its distance from the main clause; that is, a clause designated as a C clause in the process of reduction is more deeply embedded (farther from the A clause) than a B clause. Clauses have been assigned arbitrarily the following values: A = 1, B = 2, C = 3, etc. The Appendix presents the analysis of a sample text.

Applications

The clause analysis method has been used in PRDC in three different types of tests: syllogistic reasoning, arithmetic reasoning, and reading comprehension.

In the syllogistic reasoning experiment, the technique permitted the investigator to systematically control the syntactic configuration of four series of items. Two of the series, labeled "simple syntax", consisted of sentences containing no embedded clauses and two series, labeled "complex syntax" consisted of sentences containing various numbers of clauses. The syntactic configuration of all of the "simple syntax" items was identical. Individual items in the two series labeled "complex syntax" also had identical syntactic configurations (e.g., item 1 had the same number and depth of clauses in each version). The complex syntax items were constructed by matching the type of clauses (A, B, C, etc.) in each version of an item. With the number and depth of the clauses held constant, the investigator is able to study the effects of other linguistic variables on test performance — in this case the effects of vocabulary difficulty.

A method of reducing the syntactic complexity of a test of arithmetic reasoning was developed using clause analysis as a base. The purpose of the experiment was to attempt to minimize the verbal load of the test and, thus, minimize discrimination against deaf applicants who generally do not perform as well as hearing subjects on tests of arithmetic reasoning because of the high verbal content of this item type (Stunkel, 1957, for example). Preliminary results show fewer omissions on the modified items, indicating that the deaf subjects are spending less time trying to decode the verbal component of the items.

Currently we are carrying out a study to compare the efficiency and accuracy of clause analysis in describing the difficulty of reading materials with that of the Flesch Reading Ease formula. Correlations so far are in the high sixties. Clause analysis gives the item writer a coherent framework of linguistic structure in which to work. For example, an item writer trying to construct items within the reading level of a specific job would know to avoid (or maximize) the use of deeply embedded clauses. Clause analysis provides a linguistic guide to be used during the item writing process itself, as opposed to the Flesch, which serves principally as a post-facto index of reading level.

Because it is easily and quickly calculated, clause analysis would be a valuable tool in determining the reading level of jobs and in developing tests with a demonstrably job-related verbal component.

References

Bever, Thomas G. Perception, thought and language. In Roy O. Freedle & John B. Carroll (Eds.), Language comprehension and the acquisition of knowledge. Washington, D.C.: V.H. Winston & Sons, Inc., 1972.

Bloom, L.M. Language development: Form and function in emerging grammars. Cambridge, Mass.: M.I.T. Press, 1970.

Cook, Walter A., S.J. Stylistics: Measuring style complexity. Language and Linguistics Working Papers, 1975, 11, 106-120.

Flesch, R. A new readability yardstick. Journal of Applied Psychology, 1948, 32, 221-233.

Stunkel, E.R. The performance of deaf and hearing college students on verbal and nonverbal intelligence tests. American Annals of the Deaf, 1957, 102, 342-355.

APPENDIX

Clause Analysis of a Sample Text

1. #A Chomsky's position not only is unique within linguistics at the present time

2. B but is probably unprecedented in the whole history of the subject.

3. #A His first book (4) (5) revolutionized the scientific study of language

4. B published in 1957

5. B short and relativley non-technical though it was

6. +A and now he speaks with unrivaled authority on all aspects of grammatical theory.

7. #A That is not

8. B to say, of course,

9. C that all linguists, or even a majority of them, have accepted the theory of transformational grammar

10. D that Chomsky put forward some thirteen years ago in Syntactic Structures.*

The symbol "#" indicates sentence boundaries.
The symbol "+" indicates block boundaries.

Computation of Indices

1. Average Sentence Length = number of clauses - number of sentences

   10 - 3 = 3.33

2. Average Block Length = number of clauses - number of main (A) clauses

   10 - 4 = 2.50

3. Average Clause Depth = total value of clauses - number of clauses

   19 - 10 = 1.90

---

* From Lyons, John. Noam Chomsky. New York: The Viking Press, 1970.

Kerner, Susan E. & O'Mara, Francis E., US Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia. (Wed. A.M.)

AD P001339

Commanders' Assessment of Unit Effectiveness Measures

The purpose of this research was to investigate how the important area of unit effectiveness is assessed in the Army. Data was collected from a sample of senior Army commanders regarding their perceptions of existing standard Army measures of battalion effectiveness. These measures naturally classify into three groups: (1) command indicators (e.g., AWOL rates, Articles 15); (2) readiness measures (e.g., equipment rated ready, annual general inspections); and (3) the personal judgments of subordinate Army leaders. Senior Army leaders chose those measures from all of these groups which provided for them the most accurate picture of a battalion's effectiveness. It was found that military leaders not only have predetermined attitudes toward all existing effectiveness measures, but that even when this rater bias is controlled, there exists a definite preference for specific groups of measures. The command indicators were found to have the least perceived validity and utility for Army leaders, while personal judgments and readiness measures were rated significantly higher for their credibility in assessing battalion effectiveness.

COMMANDERS' ASSESSMENT OF UNIT EFFECTIVENESS MEASURES

Susan E. Kerner-Hoeg and Francis E. O'Mara

INTRODUCTION

A critical facet of successful operations by any organization is the continual monitoring of organizational performance. The information gathered through such activity is of importance for the development of realistic goals, for planning optimal strategies for achieving these goals, and for the identification and remediation of organizational deficiencies. Nowhere is organizational effectiveness measurement more vital than in the military, given the potentially disastrous consequences of misjudging national defense capabilities. Further, the estimate of aggregate military potential has broader national implications inasmuch as such estimates influence decisions in other areas of national concern, such as development of Federal budget priorities and the formulation of U.S. foreign policy. Thus, the development of means to accurately assess the strengths and deficiencies of military units is a vital concern both for the military as well as for our larger society.

Given the importance of measuring unit effectiveness, it is not surprising that the Army has traditionally monitored quantified measures of many facets of unit operations at all echelons. These measures have encompassed such disparate areas as the compilation of the maintenance status of mission-essential equipment to the tallying of chapel attendance by unit personnel. The manifest importance of unit effectiveness assessment is evidenced by the command attention paid to it and the diversity of measures employed in this assessment. Despite such attention, however, there has been a growing body of criticism regarding the accuracy and adequacy of current methods of monitoring the effectiveness of Army units.

Much of this criticism has centered around reported deficiencies in systems of unit readiness reporting (Barzily, Catalogne, and Marlow, 1980; Bowser, 1976; Robinson, 1980; Sorley, 1980; U.S. Army War College, 1976). This degree of attention is appropriate since this system constitutes the major means by which higher echelons monitor the effectiveness of Army battalions and separate companies. Even though the data from this reporting system provide major input to the development of Army contingency plans and guide high level resource allocation decisions, the consensus of opinion of those who have examined this system is that it is seriously faulted. As an example, in the Army War College study (U.S. Army War College, 1976), questionnaires measuring perceptions regarding the Unit Status Report (USR) were administered to approximately 2100 Army personnel. A full 70 percent of this sample reported that the USR does not reflect the true readiness condition of a unit. This opinion was likewise voiced in the course of interviews with over 1200 personnel conducted as another component of this same study. In addition to some technical problems in the actual computation of indices contained on the USR, this study found two major factors undermining the accuracy and credibility of this reporting system. The first factor concerned the substantial degree of latitude for subjective interpretation of unit conditions that was permitted in filling out the Unit Status Report. As an example, the Training Readiness Condition index, which constitutes a major component of the unit's readiness, is based totally upon the unit commander's subjective estimate of the number of weeks of training the unit would need to be fully ready for combat. It was the conclusion of this study that "The training portion of the USR was too subjective to be anything more than a wishful-thinking guess. The training REDCONs being reported are therefore regarded as both inflated and invalid by a sizable majority of those interviewed, particularly at company level." This opinion has also been advanced by others who have examined the validity of current unit readiness reporting procedures (Robinson, 1980; Ross, Murphy, March, and Robinson, 1979; U.S. Army Concepts Analysis Agency, 1975).

The second major problem in the USR uncovered in the Army War College study concerned a conviction by those surveyed and interviewed that there was pressure on unit commanders to portray the unit's capabilities in the best possible light even to the extent of masking genuine unit deficiencies. This problem area is one which bodes ill not only for the validity of the Unit Status Report itself, but also for the validity of any systematized quantification of unit readiness indices. A key conclusion of this study therefore was that the current Unit Status Report reflected Army units not as they actually were but rather the units as all would wish them to be.

Since the publication of that study, efforts have been made to revise and improve unit readiness reporting by increasing the reliance on objective measurements of unit conditions rather than on the subjective interpretation of the unit's capabilities. As of 1980, however, Sorley held that the Unit Status Reporting system continued to suffer major deficiencies. The most central of these deficiencies continued to be the need to separate the process of evaluating and monitoring unit effectiveness from the process by which the performance of Army officers is evaluated. Sorley believes that only by removing the responsibility for unit readiness reporting from the chain of command, which likewise evaluates the performance of the individuals who provide unit effectiveness data, can the real or perceived pressure to inflate estimates of unit effectiveness be removed.

Sorley further holds that the Unit Status Report has excluded variables which are essential to combat readiness and therefore the USR can only partially reflect the total capability of the unit. Factors such as unit cohesion and the turnover and competence of key unit personnel are those which he feels are important contributors to total unit capability but which are not now employed in estimating the unit's effectiveness. Sorley likewise suggests that the information contained on the USR be complemented with the professional judgment of individuals familiar with the unit. This, of course, would only be feasible where this judgment could be rendered frankly and openly.

The evaluation of unit effectiveness in the Army is not restricted to the USR. The Army has had a long tradition of monitoring an extensive series of variables which purportedly reflect the state of morale and discipline in the unit. Known collectively as "command indicators" or "traditional indicators," this set of unit measures typically includes such variables as reenlistment rates, crime rates, and indices associated with the administration of military justice. Unlike the USR measures, these indices are not systematically reported at the unit level to the higher echelons of the Army command structure. However, unit measures on these variables are used frequently at the local level as indicants of unit conditions and problems. Sorley (1979) has been critical of the use of such measures inasmuch as he sees them leading to a "management by statistics" in which those factors which are more readily quantifiable are given greater command emphasis than those which are more difficult to quantify, but which more substantively support and reflect unit effectiveness. Too often, he feels, command attention is expended on "getting the numbers right" in such areas of questionable military value as motor vehicle accident rates or the number of letters of indebtedness among unit personnel. This occurs at the expense of diverting command attention from such areas as unit training and equipment maintenance, which are more directly supportive of the unit's mission. The position underlying his assertions is that statistical indices of unit operations, particularly those relevant to the personnel area, are of questionable utility in assessing areas pertinent to unit effectiveness. This is somewhat inconsistent with his position (Sorley, 1980) that the USR be supplemented with measures in such areas as drug abuse, race relations, and the alienation and commitment of the unit personnel. Clearly, some statistical indices are more germane to unit effectiveness assessment than others. The question remains unresolved as to the identity of these measures. The proliferation of statistical indices used to monitor unit functioning has been fed by the variation in opinion as to which of the wide variety of possible measures are the most accurate indicants of unit capability. This proliferation has in turn led to many of the abuses and problems which have been identified in the literature.

The purpose of this research is to contribute to the resolution of these problems by examining the value of the most typically employed statistical indices in reflecting unit effectiveness. To date there has been no systematic examination across the broad spectrum of unit effectiveness measures which would permit a determination of the relative value of these measures. The absence of such information leaves unchallenged the possible reliance on inaccurate or incomplete assessment of unit effectiveness and thus the development of priorities based on apparent rather than real problems.

METHOD

Forty-eight battalion commanders, twenty-eight brigade commanders, and eight general officers located at six CONUS installations were interviewed on the topic of battalion effectiveness. During the approximately one-hour-long interviews, each subject was asked to discuss the most pressing management problems confronting him in maintaining readiness, to operationally define battalion effectiveness, and to evaluate the performance of his subordinate battalions. Each subject was also asked to assess various given measures of battalion effectiveness. These measures can be classified into three groups: Readiness Measures, Command Indicators, and Personal Judgments.

The first of these groups, Readiness Measures, is a relatively direct assessment of a unit's capability to perform its mission. This group of measures includes the REDCON ratings from the USR, the percentage of unit equipment that is operational (as gleaned from USR data) as well as ARTEP and AGI results. A listing and definition of each of these measures can be found in Figure 1.

Command Indicators are those measures which are traditionally held to reflect a unit's state of morale and discipline. The specific Command Indicators used in this study are listed and defined in Figure 2.

Personal Judgments are those opinions of battalion effectiveness held by individuals at various echelons. The ascending hierarchy of authority and the six specific levels used in this study are as follows: (1) Service members in the battalion (SM); (2) Noncommissioned officers in the battalion (NCOs); (3) Company grade officers in the battalion; (4) Brigade Commanders; (5) Assistant Division Commanders; and (6) Division Commanders.

Subjects were asked to indicate "how accurate an assessment of battalion effectiveness would be if it were based on any single piece of information from the list provided." A measure providing complete accuracy would be rated 100%, while a measure providing no information on unit effectiveness would be rated 0%.

Subjects were further asked to choose from the given list of measures the five which, in combination, would provide "the most complete picture of a battalion's overall effectiveness."

RESULTS

Analyses of these data began with an examination of the degree to which there was a difference in the perceived validity of the effectiveness measures across positions (i.e., battalion commander vs. brigade commander vs. general officer). A three-level one-way ANOVA was therefore performed on the validity ratings given to each of the unit effectiveness measures. Of the twenty-two measures tested, on only one was there a significant position difference (Drug Arrest Rate). It was concluded that there existed no consistent position differences in the perceived validity of the unit effectiveness measures, since such a proportion of significant results is essentially what would be expected from chance alone. Accordingly, the data from the three groups were combined in all further analyses.

The mean accuracy ratings assigned to each measure are rank ordered and presented in Table 1. As shown, a wide range of mean ratings was obtained, varying from 72.5% accuracy attributed to ARTEP results to an accuracy rating of only 29.2% for desertion rates. In general, the Readiness Measures and the Personal Judgments were given the highest validity as measures which individually render an accurate assessment of a battalion's effectiveness.

Table 2 presents a rank ordering of the frequency with which each measure was included in the group of five providing the most complete picture of a battalion's effectiveness. The sharp drop in the frequency of selection after the fourth measure indicates that these first four are important measures in providing an overall effectiveness assessment. Measures of readiness, specifically the ARTEP and AGI, and the personal judgments of those in the unit, specifically the company grade officers and NCOs, were seen to provide the most information about battalion effectiveness. Further, the top four measures in Table 2 are the same as the top four in Table 1, implying that these measures are the most valid, whether they are considered individually or in combination. For the most part, the group of command indicators are located at the bottom of the continuum on both Tables 1 and 2.

As seen in Table 1, all measures within each group (Command Indicators, Personal Judgments, Readiness Measures) tended to be assigned similar accuracy ratings. Thus, there appears to have been a consistent rating applied to all measures within each of the three groups of measures. That is, while the judgments of unit NCOs may have been accorded higher validity by subjects than the judgment of division commanders, the fact that both measures entail personal judgments tended to produce very similar validity ratings for both measures. To test this, coefficient alphas were computed for each of the three groups to determine the internal consistency of this grouping. These reliability measures are presented in Table 3A. All three coefficient alphas are above .89, revealing a high degree of internal consistency within each group of effectiveness measures.

In order to detect whether there was a consistent style of rating (i.e., preference of one group of measures to the exclusion of others), the relationships among the three groups of measures was examined. A mean rating was computed for each subject for each of three groups of measures. Pearson correlations were in turn computed among these mean ratings. These intergroup correlations are presented in Table 4A. A substantial positive correlation exists among the three groups, suggesting that even in the presence of a wide variation of mean ratings (as shown in Table 1), there was a tendancy for subjects to display a rater bias reflecting a global impression of the validity of any formal effectiveness measure. Thus, a subject who gave high ratings to one group of measures likewise gave high ratings to the other two groups. As an extreme example, one subject's accuracy ratings of the individual measures ranged from 90-100, while another's ranged from 0-16.

Thus, to correct for this rater bias, a set of corrected ratings was computed for each subject on each measure. This was accomplished by first computing the average rating on all measures given by each subject. Each subject's average rating was then subtracted from his original rating on each measure to establish a set of corrected ratings for each subject.

The rank ordering for these corrected ratings agrees totally with the rank ordering of the uncorrected ratings (Table 1), indicating that these earlier results were not artifacts of subjects' biases toward effectiveness measures in general. Thus, the ordering of the ratings for those with high confidence in unit effectiveness measures is the same as that for those with low confidence in unit effectiveness measures. The existence of this rater bias, however, does raise the possibility that the high internal consistencies in each of three groups of measures was not due to these existing as natural groupings in the minds of the raters, but rather was an artifact of the rater bias. It is possible that the rater bias produced high intercorrelations among the ratings given to all of the unit performance measures. This would in turn produce high coefficient alphas for the ratings given to the measures within each of the groups. This interpretation is made plausible by the substantial positive correlations among the mean ratings given to each group of measures (Table 4A). To test this interpretation, coefficient alphas were recomputed for each of the three groups, based on the corrected ratings. These reliability measures are presented in Table 3B. Though the alphas drop slightly from those based on the uncorrected ratings, the internal consistency of the groups remains acceptable, indicating that while these coefficient alphas were inflated by the rater bias, they were not totally attributable to them.

Intercorrelations among the three groups of measures were next computed, based on the corrected ratings. These intercorrelations are presented in Table 4B, showing correlations which are negative, in contrast to those based on uncorrected ratings (Table 4A). The use of difference scores to correct for the rater bias reveals a clearer picture of Army leaders' preferences for different types of measures. That is, a definite tendency to favor one type of measure over the other two is now displayed.

DISCUSSION

The results of this research indicate that the evaluation of the existing Army battalion effectiveness measures has two facets. First, there is the attitude military leaders have towards the general category of formal battalion effectiveness measures. There was seen wide variation from commander to commander in this attitude, with some commanders attributing very little credibility to any of the battalion effectiveness measures. Second, there appears to exist in the perceptions of senior commanders a distinct typology of battalion effectiveness measures. The wide spectrum of measures studied in this research broke down into only three types, or groups, of measures in the eyes of the interviewed commanders--Readiness Measures, including formal evaluations of unit capability; Personal Judgments, the estimation of battalion effectiveness by individuals within the unit and above the battalion in the chain of command; and Command Indicators, which include traditional measures of unit morale and discipline. These groups were seen to have differing degrees of utility as indicants of battalion effectiveness. Readiness Measures and Personal Judgments were seen to be similar in value whereas Command Indicators were seen to have substantially less merit as indices of unit effectiveness. These findings are very much in line with those of other studies both in military (Bowser, 1976) and in civilian organizations (Mahoney & Weitzel, 1969; Weitzel, Mahoney & Crandall, 1971) which show that managers display a clear preference for operationally oriented measures as indices of organizational performance, especially in contrast to measures related to the personnel and human relations areas.

From these results, at least two points are salient in their operational implications: the low value ascribed to Command Indicators as measures of unit effectiveness and the high value ascribed to Personal Judgments.

The first of these two results calls into question the practice of using Command Indicators as measures of unit effectiveness. This finding is surprising in light of the long history of utilization which such measures have had in the military. Despite this tradition, the interviewed commanders saw each of these measures as being ambiguous in their implications regarding unit effectiveness since a high score on any of these measures could stem from a multitude of possible causes. Thus, while each of these measures may truly reflect the status of various personnel areas in the unit, this information by itself is held to be too ambiguous to judge the effectiveness of the command. The use of such measures should therefore be restricted to the monitoring of personnel trends and issues at levels above the unit. If such measures are used at the unit level to assess overall unit effectiveness, the present results suggest that these measures should be used only in combination with other indicants of unit operation. In using Command Indicators to assess the morale and the state of discipline of a given unit, especially high or low scores on these measures should not be taken at face value but rather should precipitate a fuller investigation into the reasons behind the scores. Only with this fuller base of information can sound conclusions be reached regarding the state of the unit.

The relatively high validity ascribed to the Personal Judgment measures supports Sorley's contention that an ideal Unit Status Reporting system would include the reporting of the professional judgment of battalion effectiveness by key personnel. In light of the already well-documented pressure on unit personnel to have USR data appear maximally positive, it is doubtful that judgments rendered within the context of the present USR system would substantially contribute to a fuller assessment of unit readiness at higher command levels. At the unit level, however, the high ratings given to NCO and junior officer judgments of battalion effectiveness support the notion that unit commanders can best assess the day-to-day status of their units by relying on the input of their subordinates. It is these individuals who have the most detailed and direct knowledge of unit strengths and weaknesses inasmuch as it is they who directly address these areas in the course of their daily duties. For such a rich source of information to be used to its fullest potential, however, the commander must have the skill to solicit this information in a frank and unbiased form. The high value attributed to these estimates of unit performance, therefore, underlines the importance of leadership development for the Army. Without the requisite leadership skills to involve his or her subordinate leaders in the development of the total unit potential and to solicit and evaluate their frank judgment of unit capabilities, unit commanders would be deprived of one of the major instruments of monitoring and improving the operation of their units.

In the concept of prior work in this area (Concepts Analysis Agency, 1975; Robinson, 1980; Ross, et. al., 1979; Sorley, 1979; US Army War College, 1976), the present results call for a careful re-examination of the structure and process currently employed to evaluate unit effectiveness. There continues to be a pressing need to amend and improve current unit effectiveness assessment systems used in the military. The process is ongoing; it needs to continue and accelerate.

REFERENCES

Barzily, Z., Catalogne, P. R., & Marlow, W. H. Assessing Marine Corps Readiness (T-430). Washington, D.C.: The George Washington University, September 1980.

Bowser, S. E. Determination of Criteria of Operational Unit Effectiveness in the US Navy (NPRDC TR 76TQ-41). San Diego, California: Navy Personnel Research and Development Center, August 1976.

Mahoney, T. A. & Weitzel, W. Managerial Models of Organizational Effectiveness. Administrative Science Quarterly, 1969, 14, 357-365.

Robinson, R. M. Objective Measurement of Training Readiness. Carlisle Barracks, Pennsylvania: US Army War College, May 1980.

Ross, G., Murphy, J., March, M., Robinson, R., & Tullington, B. Military Organizational Effectiveness/Readiness and Sustainability. Final report to Headquarters, Department of the Army, under Contract DAAK21-79-C-0015, McLean, Virginia: Science Applications, Inc., September 1979.

Sorley, L. Professional Evaluation and Combat Readiness. Military Review, October 1979, 41-53.

Sorley, L. Prevailing Criteria: A Critique. In Sarkesian, S. C. Combat Effectiveness--Cohesion, Stress, and the Volunteer Military. Beverly Hills: Sage Publications, 1980.

US Army Concepts Analysis Agency, Readiness System Study, Phase I Analysis. AD-A029 387, August 1976.

US Army War College, US Army Unit Readiness Reporting. ACN 75025, June 1976.

Weitzel, W., Mahoney, T., & Crandall, N. A Supervisory View of Unit Effectiveness. Final report to Office of Naval Research, under Contract ONR N00014-68-A-0141-0003, Minneapolis, Minnesota: Industrial Relations Center, University of Minnesota, 1971.

Figure 1

READINESS MEASURES

OVERALL READINESS	A battalion's overall readiness status as reported in the monthly Unit Status Report.
PERSONNEL READINESS	A battalion's personnel readiness status as reported in the monthly Unit Status Report.
EQUIPMENT ON HAND	An index of the degree to which a battalion possesses all authorized equipment, a reflection of the battalion's supply system.
EQUIPMENT SERVICEABILITY	The maintenance status of a battalion's equipment, a reflection of the battalion's maintenance system.
EQUIPMENT ON HAND RATED READY	The proportion of equipment a battalion actually has on hand that is operational.
ARTEP	The percentage of the missions/tasks rated "satisfactory" during a battalion's most recent field training exercise.
AGI	The percentage of the areas rated "satisfactory" during a battalion's most recent annual general inspection.

Figure 2

COMMAND INDICATORS

ARTICLES 15	The percentage of enlisted personnel administered non-judicial punishment (e.g., fines, reductions in grade) during a given month.
COURTS MARTIAL	The percentage of enlisted personnel receiving a court martial during a given month.
AWOL	The percentage of enlisted personnel who were involved in unexcused absences during a given month.
DESERTIONS	The percentage of enlisted personnel who deserted during a given month.
FIRST TERM RE-UP	The percentage of a battalion's first-term reenlistment objective that was achieved in a given month.
CAREER RE-UP	The percentage of a battalion's reenlistment objective for career personnel that was achieved in a given month.
CRIMES OF VIOLENCE	The percentage of a battalion's enlisted strength involved in crimes of violence in a given month.
PROPERTY CRIMES	The percentage of a battalion's enlisted strength involved in crimes against property in a given month.
DRUG ARRESTS	The percentage of a battalion's enlisted strength arrested for drug and marijuana violations in a given month.

Table 1

Accuracy Assessments of The Effectiveness Measures

Measure	Mean Accuracy Rating
ARTEP	72.5 %
Company grade officer's judgment	71.2 %
NCO's judgment	71.0 %
AGI	66.7 %
Equipment on hand rated ready	63.3 %
Brigade commander's judgment	62.5 %
Service member's judgment	62.1 %
Equipment status	59.9 %
Overall readiness	57.2 %
Assistant Division Commander's judgment	56.9 %
Division Commander's judgment	53.7 %
First-term reenlistment rate	51.3 %
Personnel readiness	50.7 %
Equipment on hand	46.6 %
AWOL rate	45.9 %
Career reenlistment rate	43.8 %
Crimes against property	38.1 %
Article 15s	37.4 %
Crimes of violence	36.7 %
Courts-martial rate	32.5 %
Drug/marijuana convictions	30.8 %
Desertion rate	29.2 %

Table 2

Group of Five Measures
Providing Most Complete Picture of Effectiveness

Measure	Number Choosing Measure	
ARTEP	63	(75 %)
NCO's judgment	54	(64%)
AGI	48	(57%)
Company grade officer's judgment	44	(52%)
Brigade commander's judgment	28	(33%)
First-term reenlistment	27	(32%)
Service member's judgment	25	(29%)
Overall readiness rating	25	(29%)
Equipment on hand rated ready	21	(25%)
AWOL rate	15	(18%)
Personnel readiness	13	(15%)
Equipment status	12	(14%)
Article 15s	9	(11%)
Career reenlistment rate	7	(08%)
Division commander's judgment	7	(08%)
Assistant division commander's judgment	5	(06%)
Crimes against property	4	(05%)
Drug/marijuana convictions	3	(04%)
Crimes of violence	3	(04%)
Equipment on hand	3	(04%)
Courts-martial rate	2	(02%)
Desertion rate	1	(01%)

Table 3

Reliability Coefficients of Effectiveness Measure Groups

	A Uncorrected Ratings	B Corrected Ratings
Readiness Measures	.914	.656
Command Indicators	.950	.704
Personal Judgments	.894	.644

Table 4

Effectiveness Measure Intergroup Correlations

	A Uncorrected Ratings			B Corrected Ratings		
	Readiness Measures	Command Indicators	Personal Judgments	Readiness Measures	Command Indicators	Personal Judgments
Readiness Measures	1.000			1.000		
Command Indicators	.6537	1.000		-.6641	1.000	
Personal Judgments	.6297	.6802	1.000	-.3391	-.4781	1.000

Kimmel, Melvin J. & O'Mara, Francis E., US Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia. (Wed. A.M.)

The Measurement of Morale

An instrument measuring organizational morale was constructed from unit member satisfaction response aggregated to the battalion level. The data was gathered at three different points in time from military personnel within 55 CONUS battalions. Significant positive correlations between the satisfaction scores and an independent index of affective orientation supported the widely held, but rarely tested assumption that satisfaction measures are a true indicant of an individual's affective orientation toward his/her unit. Analysis of the instrument's psychometric properties showed it to be a reliable and valid measure of morale as an organizational characteristic as distinct from an individual level variable. Theoretical and applied implications of these findings for the study of organizational morale in military and nonmilitary units are discussed.

THE MEASUREMENT OF MORALE

Melvin J. Kimmel and Francis E. O'Mara

US Army Research Institute for the Behavioral and Social Sciences

While there are obviously many factors contributing to mission accomplishment, one that has been consistently emphasized by military strategists is the unit's morale. In a recent review of the morale literature, Motowidlo et al (1976) concluded that "apparently, hardly any military Commander doubts that morale is a potent force determining group effectiveness." (p. 52). However, these authors also point out that despite its stated importance, no coherent theories of organizational morale exist and there is virtually no systematic empirical literature on the subject.

An important first step to learning about morale would be to construct a reliable and valid measure of the concept. Motowidlo and Borman (1977) were only partially successful in developing such an instrument. The authors report some evidence for the scale's convergent validity. However, its reliability was low and there were indications of halo error in ratings. The major purpose of the present study was to develop an instrument free from such deficiencies.

One obvious issue to consider before developing a valid morale measure is its definition. Unfortunately there are almost as many definitions of morale as there are people writing about it (Motowidlo et al, 1976). While definitions differ, most writers (e.g. Guion, 1958; Martin, 1965; Stagner, 1958) seem to agree that morale represents an affective orientation toward the work unit or organization and includes "job satisfaction" as one of its major components. It would therefore appear appropriate to aggregate member responses to a series of job satisfaction items to obtain an affective measure of the unit's morale.

However, some organizational psychologists have questioned such an approach. Blum and Naylor (1968) and Motowidlo et al (1976), for example, argue that an adequate definition of morale should include such factors as motivation and cohesion, and not be limited to job satisfaction alone. Further, Guion (1973) and Lincoln and Zeitz (1980) contend that while it is possible to aggregate scores on an individual level variable to form an organizational attribute, it makes little sense to do so with an affective characteristic such as satisfaction. They explain that satisfaction, like all evaluative or affective constructs, is subject to an individual's unique motives, values and job environment. Since these characteristics differ from individual to individual, they believe it would be pointless to aggregate satisfaction scores in an attempt to form a relatively stable and generally agreed upon affective orientation toward the organization. This assumption will be tested as part of our attempt to develop a reliable and valid organizational measure of morale.

The development of this morale measure proceeded in two phases. The first involved the comparison of satisfaction scores against a derived index of affect to assess the validity of using individual satisfaction measures to represent a member's affective orientation toward the organization. The second phase was directed at examining the psychometric properties of a unit morale measure that is based on aggregated satisfaction scores.

Method

Subjects. Data was collected at three different points in time from a sample of 55 combat arms, combat support, and combat service support battalions located at six CONUS installations. At each wave of data collection an independent sample of service members, NCOs, and officers within each unit was randomly drawn, using the last digit of individual social security numbers. The total sample for each wave consisted of 6,979 service members, 5,882 NCOs and 6,172 officers.

Procedure and measures. Satisfaction measures were administered on three separate occasions at six-month intervals to the sample of unit personnel as part of a larger climate survey instrument. Surveys were administered in large groups by teams of researchers using standardized instructions. These satisfaction measures were drawn from the Survey of Organizations (Taylor and Bowers, 1972) and measured individual satisfaction with his or her unit, supervisor, coworkers, and job. The four areas of unit, supervisor, coworkers and jobs likewise defined the four major content domains of the overall climate survey. Subjects responded to each of these items utilizing a five-point scale ranging from 1 ("Strongly Disagree") to 5 ("Strongly Agree").

A morale score for each battalion was generated by first averaging the battalion members' responses to the four satisfaction items into a "General Satisfaction" score for each individual. The General Satisfaction scores for all battalion members were then averaged to derive the battalion morale measure.

The independent index of affect, used to determine if satisfaction represents a member's affective orientation toward the organization, was constructed by first converting all item responses on the climate survey to standard scores. All non-satisfaction items were next categorized as being affectively positive, negative, or neutral by two independent judges. The z-scores for all positively and all negatively rated items were then averaged separately, while the neutral items were eliminated from further analysis. The two resulting statistics were labeled z_+ and z_-, with the first of these being an indicant of a subject's tendency to agree to affectively positive items and the second reflecting agreement to affectively negative items. Highly significant ($p < .001$) negative correlations were observed between z_+ and z_- of -.50, -.49, and -.47, for waves 1, 2, and 3 respectively, suggesting that subjects were selectively attending to the affective content of the items and were responding in a manner consistent with their generalized affect towards their situations. A single index of this affective orientation ($\bar{z}$) was then produced using the equation: $\bar{z} = (z_+) - (z_-)/2$.

Results

The validity of satisfaction measures as indicants of affective orientation was determined by its relationship with the independent measure of affect, $\bar{z}$. Table 1 shows the correlations between $\bar{z}$ and the satisfaction measures taken individually and as a group for each of three waves of data collection. It is clear that there is substantial correlation between the satisfaction measures and the independent measure of affective orientation, thus validating the hypothesis.

Given a high degree of intercorrelation among the four satisfaction items, responses to these four items were averaged to produce a single General Satisfaction score for each individual. This single measure of General Satisfaction was then employed in examining the validity of affective orientation as an organizational attribute.

Two different approaches were employed in assessing this validity. The first approach examined the discriminant validity of the General Satisfaction measure at the battalion level. If General Satisfaction varied only at the individual level, it would be randomly distributed across battalions such that battalions would not differ on this variable. However, if affective orientation was a true organizational attribute, then different battalion settings would produce different levels of the General Satisfaction variable. Accordingly, the 55 battalions were compared on General Satisfaction using a least-squares one-way ANOVA on this measure. As shown in Table 2, battalions differed significantly on this measure at each rank level, and this finding was consistent across the three waves. This suggests that affective orientation is a true organizational attribute and can thus be analyzed at this level.

The second approach was to determine the stability of satisfaction at the organizational level. Battalion morale would be expected to vary somewhat from one time period to another due to differences in environmental conditions and the high level of personnel turnover within units. However, if morale is, in fact, a true organizational variable, some consistency should be observed across time, and one would expect positive correlations in battalion morale across the six months separating the data collection waves. To test this hypothesis, the battalion members' General Satisfaction scores were aggregated at each wave to produce a mean battalion morale score. Correlation coefficients were then computed between morale scores on the adjacent waves for each rank group separately. Table 3 presents the results of this analysis. As can be seen, the correlation coefficients between the Wave 1 and Wave 2 morale scores were significant at each rank level, offering some additional support for the hypothesis that morale can be conceptualized as an organizational variable. However, the hypothesis was not totally confirmed as only service members showed a significant relationship in the Wave 2/Wave 3 comparisons.

Discussion

The significant correlations found between the satisfaction items and our independent measure of affect support the commonly held, but largely untested, assumption that job satisfaction directly reflects an individual's affective orientation toward his/her work environment. More importantly, the findings derived through the analyses of the morale measures take affect out of the realm of individual psychology and suggest that an affective variable such as morale can be legitimately operationalized at the organizational level. This conclusion is tempered somewhat by the finding that the morale measure was not as stable for NCOs and officers as it was for service members. One possible interpretation is that a different set of dynamics operate upon morale at the higher levels. Another possibility relates to the fact that each aggregate score at the higher rank levels was based on a smaller _N_ than that derived for service members. This being the case, the NCO and officer morale measures would not possess the same degree of reliability as the service member data, and may underlie the attenuated stability of results at the higher levels. Further research is needed to clarify this issue.

The general conclusion that morale is an organizational variable does not contradict organizational psychologists like Guion (1973) who note the importance of distinguishing between attributes of people and attributes of organizations. However, in support of Lincoln and Zeitz (1980), the results clearly demonstrate that it is possible to obtain a relatively stable and generally agreed upon organizational measure by aggregating individual-level variable scores. While supporting the general proposition advanced by Lincoln and Zeitz, the results refute their assertion that affectively-laden concepts such as job satisfaction should not be aggregated. In making this assertion, Lincoln and Zeitz, like Guion (1973), appear to be inappropriately equating affective orientation and job satisfaction. The fact that satisfaction is an individual-level variable in no way implies that all forms of affect must be conceptualized at this level. Some characteristics, like morale, can be viewed as shared attributes of group members and, hence, qualify as organizational variables. We suggest the proper level at which to conceptualize and operationalize a construct should be empirically determined rather than decided upon on an a priori basis.

A separate question relates to the adequacy of a morale measure that is based solely on satisfaction. Although most writers agree that satisfaction is an important component of morale, some, like Blum and Naylor (1968) and Motowidlo and Borman (1977) argue that other dimensions should also be included to capture its full meaning. We suggest that while morale may, in fact, be a multidimensional variable, this does not necessarily imply that a unidimensional measure such as the one described in the present paper is inappropriate. Operational definitions of psychological constructs rarely (if ever) tap all relevant dimensions. The field of psychology has usually progressed by beginning with limited measures of a particular concept and subsequently building upon these first approximations (Elms, 1975). The same procedure is suggested in the case of organizational morale. Other hypothesized dimensions should be incorporated into future measurement instruments and their discriminent and concurrent validity tested. This systematic approach should lead to the development of a truly reliable and valid instrument that does justice to the potential multidimensional nature of the concept. Once adequate measures are developed, it will then be possible for researchers to effectively study the antecedents and consequences of the morale construct, which is thought to play such an important role in the life of a military organization.

REFERENCES

Blum, M. L. and Naylor, J. C. Industrial Psychology: Its Theoretical and Social Foundations. New York: Harper & Row, 1969.

Elms, A. C. The crisis of confidence in social psychology. American Psychologist, 1975, 30, 967-76.

Guion, R. M. Industrial morale (a symposium) I: the problem of terminology. Personnel Psychology, 1958, 11, 59-64.

Guion, R. M. A note on Organizational Climate. Organizational Behavior and Human Performance, 1973, 9, 120-123.

Lincoln, J. R. and Zeitz, G. Organizational properties from aggregate data: Separating individual and structural effects. American Sociological Review, 1980, 45, 391-408.

Martin, A. R. Morale and productivity: a review of the literature. Public Personnel Review, 1965, 30(1), 42-45.

Motowidlo, S. J., and Borman, W. C. Behaviorally anchored scales for measuring morale in military units. Journal of Applied Psychology, 1977, 62(2), 177-183.

Motowidlo, S. J. Dowell, B. E., Hopp, M. A., Borman, W. C., Johnson, P. D. and Dunnette, M. D. Motivation, Satisfaction, and Morale in Army Careers: A Review of Theory and Measurement. Minneapolis: Personnel Decisions, Inc., 1976.

Stagner, R. Industrial morale (a symposium) II: Motivational aspects of industrial morale. Personnel Psychology, 1958, 11, 64-70.

Taylor, J. C. and Bowers, D. G. Survey of Organizations. Ann Arbor: Institute for Social Research, 1972.

Table 1. Zero-order and Multiple Correlations between Affective Response Bias ($\bar{z}$) and Satisfaction Measures

	Wave 1		Wave 2		Wave 3	
	r	R	r	R	r	[illegible]
Satisfaction with Job	.61		.62		.60	
Satisfaction with Unit	.63		.63		.63	
Satisfaction with Supervisor	.65		.63		.65	
Satisfaction with Coworkers	.44		.46		.44	
		.81		.79		.8[illegible]

Table 2. Results of One-way ANOVAs Testing Discriminability of General Satisfaction Measure by Rank and Time

	Time		
Rank Level	1	2	3
	F (df)	F (df)	F (df)
Service Members	2.748 (53,3577)	3.649 (52,3789)	3.055 (53,4040)
NCOs	2.902 (52,1566)	2.689 (52,1517)	1.794 (53,1755)
Offices	1.908 (48,460)	1.874 (52,597)	3.076 (53,645)

NOTE: $p < .001$ for all F valves.

Table 3. Morale Measure Intercorrelations across Adjacent Waves by Rank Level

Rank Level	Interwave Coefficients r Wave 1/Wave 2	Interwave Coefficients r Wave 2/Wave 3
Service Members	.3881**	.4109**
Noncommissioned Officers	.2463*	.1941
Officers	.4945**	.1755

** $p < .01$
* $p < .05$

Early identification of Academic Difficulties in Foreign Language Training

MAJ Allen N. Knox
Deputy Chief, Test Division
Defense Language Institute
Foreign Language Center

DLIFLC utilizes multiple sequential screening techniques for first-term enlistees of all Services who apply for foreign language training. Candidates must qualify on the AFQT, attain specified scores on selected parts of the ASVAB and attain a specified minimum score on the Defense Language Aptitude Battery. Despite these measures, academic attrition ranges from 12% to 30%, depending upon the language studied. A 56-item, five-option, attitude/interest instrument was constructed and administered to students before their courses started. Item leads centered on study habits, career goals, intercultural relationships and personal attributes. The five options varied from "strongly agree" to "strongly disagree." A scoring program permitted assignment of varying weights (4 to 0) to item options. For analysis, students were divided among three achievement groups based on 6th-week grades. This paper discusses the use of cross-tabulation tables (item option by achievement group) for item analysis and various option weighting strategies used to improve discrimination among achievement groups.

The purpose of this research was to develop a questionnaire which would complement the Defense Language Aptitude Battery in predicting those students of foreign language likely to encounter academic difficulty early in their course of instruction. The purpose of this paper is to describe some problems encountered in using cross-tabulation tables to weight item options to maximize predictive validity of the questionnaire.

Introduction: Enlisted students at the Defense Language Institute, Foreign Language Center (DLI) take at least three pencil-and-paper aptitude tests before being selected for attendance; Armed Forces Qualification Test, Armed Services Vocational Aptitude Battery (ASVAB), and Defense Language Aptitude Battery (DLAB). These tests and the nature of a volunteer military tend to cause the enlisted student body to be relatively homogeneous with respect to age and educational experience.

Despite selection procedures, academic attrition is approximately 14% and varies with the difficulty of the language being learned. Another predictive instrument (hereafter referred to as FLII) was deemed desirable to enable management to either (1) reduce attrition by applying different teaching strategies to those students whose DLAB and FLII scores indicated probable difficulty, or (2) to release early in the course those students whose DLAB, FLII and course scores indicated academic difficulty and probable attrition.

Management requirements for the FLII were that its development consume minimum resources, that it have a short administrative time (30 minutes or less) and that it be machine scoreable. Salient psychometric considerations were predictive validity and low correlation with DLAB.

Method: In order to meet the requirement of using minimum resources, it was decided to use items from past research projects. One source was the Foreign Language Interest Inventory developed at DLI in 1969. Because this instrument was designed to be administered to students during their course of instruction, many items were inappropriate for the new instrument or needed revision. A second source was a HumRRO study (Fiks and Brown, 1969) which had developed some items that correlated with academic success in learning foreign languages. The third source was a panel of experts who had a combined experience level in foreign language education and testing of 40 years. These three sources provided 56 items for the pilot form. Each five-option item described a factor (career goal, experience, intercultural relationships, personal attribute, etc.) thought to be related to learning a foreign language. The options typically took the form, agree strongly...not sure...disagree strongly, and the order of presentation was reversed on some items.

Items from the early Foreign Language Interest Inventory and the HumRRO instrument were originally assigned option weights of 1 (for one option) and 0 (for the remaining options). Traditional item analysis provided insight into which should be assigned 1.

Assigning option weights in this manner seemed undesirable for items purportedly measuring a continuum. A great deal of information may be lost when one option is weighted as "correct" and all others "incorrect". Likewise, for statistical analysis, a great deal of variability may be lost using this weighting scheme. Therefor, a scoring program was written for the FLII which permitted assigning a weight of 0 to 4 to each item option.

The FLII was administered to 279 students before they started their course of instruction. These students were randomly divided into two groups; a derivation sample and a validation sample. Due to administrative attrition and incomplete answer sheets, N=130 and 124 for the derivation and validation samples, respectively.

Course grade at the end of six weeks was used as the criterion. Students were divided into three academic groups according to course grade: >90=Hi, 80-90=Med and <80=Lo. Table 1. summarizes information on sample size and academic group.

TOTAL SAMPLE N=254

DERIVATION N=130

ACADEMIC GROUP	HI	MED	LOW
GRADES	91-100	80-90	0-79
N=	53	50	27

VALIDATION N=124

ACADEMIC GROUP	HI	MED	LOW
GRADES	91-100	80-90	0-79
N=	53	40	31

TABLE 1. SUMMARY OF SAMPLE

Item analysis was accomplished using 3 x 5 contingency tables (academic group by item option) generated for each item by Subprogram Crosstabs from the Statistical Package for the Social Sciences (Nie, et al, 1975). Fig. 1. shows "ideal" column percentages for an imaginary item measuring a continuous variable using five options. In this ideal example, half of the high (Hi) academic group chose option A. Other options were chosen by fewer members of the high group. The students doing relatively poorly in the course (Lo) most often chose options representing the opposite end of the continuum. The middle achievers (Med) most often selected options representing the middle of the continuum.

ACADEMIC GROUP

OPTION		LOW	MED	HI
	A	0	5	50
	B	10	20	25
	C	15	50	15
	D	25	20	10
	E	50	5	0

Fig. 1. Contingency Table Showing

"Ideal" Results-Column Percentages Only

Guttman (1941) showed that, for a quantitative criterion, the criterion mean of the examinees who chose that option could be used as the option weight. However, it was decided to use Subprogram Crosstabs because it was readily available and the output was in a form useful to other research.

Results: An a priori weight table, with options scored 0, 1, 2, 3, 4 or 4, 3, 2, 1, 0 depending upon how the options were presented, was used so that a high FLII score would be indicative of high academic grades. Using this straightforward weighting scheme, the FLII score correlated (Pearson) .2966 with sixth-week grades. In light of Cronbach's (1970) assertion that "Correlations of interests with grades in related fields are generally below 0.30...", the magnitude of this correlation was encouraging.

Contingency tables for each item were analyzed. The analysis indicated that "Lo students" and "Hi students" displayed rather similar option selection patterns. Unlike the pattern in Fig. 1., differences between Hi and Lo groups were a matter of degree rather than direction. Fig. 2. shows the contingency table (column percentages only) for item 37. Whereas 84% of the Lo group and 78% of the Hi group picked options A and B, 3 of 4 "Hi students" chose B and only 1 of 2 "Lo students" picked B.

	LOW	MED	HI	ORIGINAL WEIGHT	REVISED WEIGHT
A	33	28	6	4	0
B	51	56	73	3	4
C	11	16	20	2	3
D	4	0	0	1	0
E	0	0	0	0	0

Fig. 2. Item 37 Contingency Table, Column Percentages Only. Original and Revised Weights Shown at Right.

A similar pattern of responses was found for many items, and option weights were revised following these rules:

1) A high weight (usually 4) was assigned to that option chosen most often by the Hi group.

2) Lower weights were assigned to options when more Hi than Lo selected the option.

3) Zero was assigned to those options when more Lo than Hi selecte it.

Note that no items were discarded at this time. Fig. 2. shows original and revised weights for item 37 based on this weighting strategy. FLII scores now correlated (Pearson) .09 with sixth-week grades.

Using five weights was notably unsuccessful. To reduce complex and because responses of Hi and Lo groups appeared similar to warra five weights, a simplified weighting strategy was used:

1) A weight of 2 was assigned to options where the percentage of H selecting an option exceeded the percentage of Lo by 20 or more.

2) A weight of 1 was assigned to options where the percentage difference was 10-19.

3) Zero was assigned to all other options.

The correlation between FLII score (using the new weighting) an sixth-week grades was .13. It rose to .24 when the Med group was excluded from calculation.

Since the Med group appeared to adversely affect correlation and had been ignored in weighting strategies, it was decided to include them in the weighting strategy. Fig. 1., depicting an "ideal" response pattern, shows that a pattern exists, not only within columns, but across rows.

In order to capitalize on the row and column pattern a new weight strategy was used:

1) Options A and E were assigned 4 or 0 if they closely matched the "ideal" row and column pattern. They were assigned 3 or 1 if they matched either row or column pattern.

2) Options B and D were assigned 3 or 1 if they closely matched the ideal row and column patterns or 2 if they matched only one pattern.

3) Option C weight was assigned dependent upon the pattern and weights of other options.

With this weight strategy, 7 items were discarded because of lack of any pattern. FLII score correlated .15 with sixth-week grades.

Thus far, all weighting strategies were inferior to the original, a priori weight table. It was decided to revert to the original weights, but refine them by assigning zero weight to those options chosen by 10% or less of the Hi group (See Fig. 3.). Seven items were deleted because response patterns showed no distinction between Hi and Lo groups.

	LOW	MED	HI	ORIGINAL WEIGHT	REVISED WEIGHT
A	74	77	83	4	4
B	18	18	13	3	3
C	7	4	3	2	0
D	0	0	0	1	0
E	0	0	0	0	0

Fig. 3. Item 24 Contingency Table, Column Percentages Only. Original and Revised Weights Shown at Right.

FLII scores now correlated .39 with sixth-week grades and .15 with DLAB scores. The correlation with grades regressed to .22 when the weight table was applied to the validation sample.

Conclusion: Gage (1957) found that partially a priori scores for five-option items were at least as valid for prediction as an elaborate weighting method. This research is partially supportive of those findings. While some improvement in predictive validity was possible using contingency tables for item analysis and variable option weighting, the method was trial-and-error and more time consuming than expected. The Guttman method is more rigorous and writing a computer program for its use does not appear difficult. DLI is presently writing such a program, and the results will permit a comparison between Guttman's method and the contingency table method used in this research.

References:

Cronbach, Lee J. Essentials of Psychological Testing, 3d Ed. New York: Harper & Row, 1970

Fiks, Alfred I. and George H. Brown. Student Attitudes and Foreign Language Learning. The George Washington University Human Resources Research Office, March 1969.

Gage, N. L. Logical vs. empirical scoring keys: the case of the MTAI. Journal of Educational Psychology, 1957, 48, 213-216.

Guttman, L. Supplementary Study B, pp. 251-364, in P. Horst (Ed), The prediction of personnel adjustment. New York: Social Science Research Council, 1941

Nie, Norman H., C. Hadlai Hull, Jean G. Jenkins, Karen Steinbrenner and Dale H. Bent. Statistical Package for the Social Sciences, Second Ed., New York; McGraw-Hill Book Company, 1975

COMPUTER-BASED SKILL QUALIFICATION TESTING: DEVELOPMENT AND VALIDATION

Christopher G. Koch
Judith A. Englert
Richard E. Vestewig
Honeywell Systems and Research Center
2600 Ridgway Pkwy., P.O. Box 312, Minneapolis, MN 55440

John T. Larson
U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Ave., Alexandria, VA 22333

An alternative to present SQTs for tactical data systems is the concept of a computer-based embedded SQT administered on the operational system or simulator to accomplish standardized, objective, hands-on skill qualification assessment. This paper describes a computer-based SQT Hands-On Component developed for TACFIRE Artillery Control Console Operators, MOS 13C, using the PLANIT courseware authoring language. The SQT complies with TRADOC policy and procedures for SQT development and validation (TRADOC Reg 351-2, Pam 351-2). An expert tryout for content validity and a representative-soldier tryout for feasibility of administration were conducted. A method for diagnosing and correcting logic errors in the courseware (vs. operators' errors) was applied. Benefits of computer-based SQT testing include: maximum fidelity attained by utilizing the operational system; control, monitoring, and scoring performed by the tactical system computer; centralized or field unit testing capability; immediate, detailed performance feedback provided; test results compatible in format with TRADOC guidelines.

COMPUTER-BASED SKILL QUALIFICATION TESTING: DEVELOPMENT AND VALIDATION[1]

Christopher G. Koch
Judith A. Englert
Richard E. Vestewig
Honeywell Systems and Research Center
2600 Ridgway Pkwy., P.O. Box 312, Minneapolis, MN 55440

John T. Larson
U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Ave., Alexandria, VA 22333

1.0 INTRODUCTION

1.1 Statement of the Problem

The Skill Qualification Test (SQT) is a primary means the Army uses to assess soldiers' proficiency in performing critical tasks of their military occupational specialties (MOSs). It contains performance-oriented and criterion-referenced tests on critical tasks selected from the soldier's manual for a duty position. Test results feedback is provided to soldiers, supervisors, training managers, and training/test developers. The SQT usually contains a written Skill Component (SC), a Hands-On Component (HOC), and a Job Site Component (JSC).

Although the SQT program was designed to maximize hands-on performance testing, problems of standardization, inter-rater scoring reliability, extensive time and resource requirements, training of test administrators, and overall administrative feasibility have attenuated its success. Much skill qualification testing has become primarily written in nature, via the SC or a written Alternate HOC. The written tests allow for standardized administration and scoring, but often have low fidelity to actual job performance requirements. The intended objective of performance-based, criterion-referenced SQT testing could be more effectively realized by a) increased fidelity of the SQT (that is, more similarity) to job requirements and task performance; b) reduced reliance on written components; c) reduced time between test administration and test results; and d) increased feasibility of unit level SQT administration.

1.2 Concept of Computer-Based Embedded Testing

The increasingly extensive use of the computer in military systems offers an opportunity to alleviate present shortcomings common to conventional SQTs for tactical data and weapons systems. The concept is computer-based embedded SQT testing in which the computer of a tactical weapons system monitors, controls, and scores a test on actual job tasks administered on that system itself (or a simulator). This paper describes the development and validation of a computer-based SQT Hands-On Component (HOC) for the TACFIRE Artillery Control Console Operator (ACCO), MOS 13C.[2]

---

[1] The views expressed in this paper are those of the authors and do not necessarily imply endorsement of the Department of the Army or the Army Research Institute.

[2] Work described in this paper was performed by Honeywell Inc. for the U.S. Army Research Institute for the Behavioral and Social Sciences, contract no. MDA903-78-C-0386.

COMPUTER-BASED SKILL QUALIFICATION TESTING: DEVELOPMENT AND VALIDATION

Christopher G. Koch
Judith A. Englert
Richard E. Vestewig
Honeywell Systems and Research Center
2600 Ridgway Pkwy., P.O. Box 312, Minneapolis, MN 55440

John T. Larson
U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Ave., Alexandria, VA 22333

An alternative to present SQTs for tactical data systems is the concept of a computer-based embedded SQT administered on the operational system or simulator to accomplish standardized, objective, hands-on skill qualification assessment. This paper describes a computer-based SQT Hands-On Component developed for TACFIRE Artillery Control Console Operators, MOS 13C, using the PLANIT courseware authoring language. The SQT complies with TRADOC policy and procedures for SQT development and validation (TRADOC Reg 351-2, Pam 351-2). An expert tryout for content validity and a representative-soldier tryout for feasibility of administration were conducted. A method for diagnosing and correcting logic errors in the courseware (vs. operators' errors) was applied. Benefits of computer-based SQT testing include: maximum fidelity attained by utilizing the operational system; control, monitoring, and scoring performed by the tactical system computer; centralized or field unit testing capability; immediate, detailed performance feedback provided; test results compatible in format with TRADOC guidelines.

COMPUTER-BASED SKILL QUALIFICATION TESTING: DEVELOPMENT AND VALIDATION[1]

Christopher G. Koch
Judith A. Englert
Richard E. Vestewig
Honeywell Systems and Research Center
2600 Ridgway Pkwy., P.O. Box 312, Minneapolis, MN 55440

John T. Larson
U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Ave., Alexandria, VA 22333

1.0 INTRODUCTION

1.1 Statement of the Problem

The Skill Qualification Test (SQT) is a primary means the Army uses to assess soldiers' proficiency in performing critical tasks of their military occupational specialties (MOSs). It contains performance-oriented and criterion-referenced tests on critical tasks selected from the soldier's manual for a duty position. Test results feedback is provided to soldiers, supervisors, training managers, and training/test developers. The SQT usually contains a written Skill Component (SC), a Hands-On Component (HOC), and a Job Site Component (JSC).

Although the SQT program was designed to maximize hands-on performance testing, problems of standardization, inter-rater scoring reliability, extensive time and resource requirements, training of test administrators, and overall administrative feasibility have attenuated its success. Much skill qualification testing has become primarily written in nature, via the SC or a written Alternate HOC. The written tests allow for standardized administration and scoring, but often have low fidelity to actual job performance requirements. The intended objective of performance-based, criterion-referenced SQT testing could be more effectively realized by a) increased fidelity of the SQT (that is, more similarity) to job requirements and task performance; b) reduced reliance on written components; c) reduced time between test administration and test results; and d) increased feasibility of unit level SQT administration.

1.2 Concept of Computer-Based Embedded Testing

The increasingly extensive use of the computer in military systems offers an opportunity to alleviate present shortcomings common to conventional SQTs for tactical data and weapons systems. The concept is computer-based embedded SQT testing in which the computer of a tactical weapons system monitors, controls, and scores a test on actual job tasks administered on that system itself (or a simulator). This paper describes the development and validation of a computer-based SQT Hands-On Component (HOC) for the TACFIRE Artillery Control Console Operator (ACCO), MOS 13C.[2]

---

[1] The views expressed in this paper are those of the authors and do not necessarily imply endorsement of the Department of the Army or the Army Research Institute.

[2] Work described in this paper was performed by Honeywell Inc. for the U.S. Army Research Institute for the Behavioral and Social Sciences, contract no. MDA903-79-C-0386.

The SQT HOC is comprised of courseware written in a computer-assisted instruction system and authoring language called PLANIT. It is administered in an embedded mode on the TACFIRE tactical system either in the school or field setting. The SQT complies with TRADOC policy and procedures for SQT development and validation (TRADOC Reg 351-2, Pam 351-2) and offers many advantages over conventional HOC testing for the MOS 13C duty position.

1.3 Background

1.3.1 TACFIRE System. The TACFIRE system selected as the testbed for applying the embedded SQT concept is a computerized TACtical FIRE direction system designed to coordinate the command, control, and communications functions within and among all levels of the field artillery system from corps to forward observers. TACFIRE may be conceptualized as an interrelated set of subscribers, each of whom has access to a common data base and a powerful set of applications software. The artillery control console (ACC) shown in Figure 1 is the principal man-machine interface in the TACFIRE system.

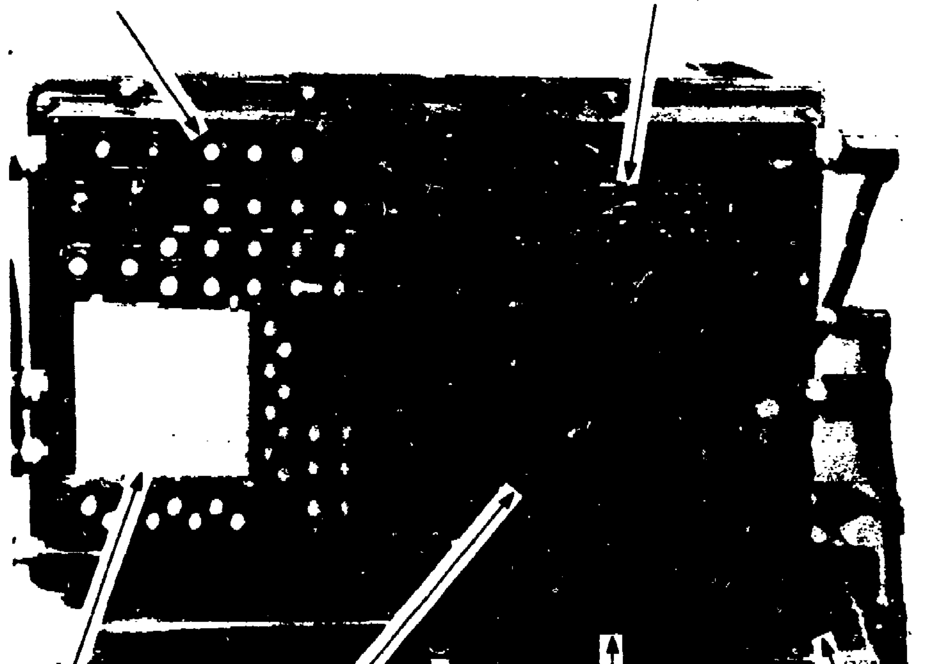


Figure 1. TACFIRE Artillery Control Console

The ACCO is a grade E6 or E7 noncommissioned officer responsible for maintaining and updating the TACFIRE data base. His functions include: entering data required for artillery target intelligence, fire planning, tactical and technical fire control, and commander's criteria; controlling communications among the network of TACFI subscribers; and initiating, processing, and/or terminating artillery fire missions. Task activities for which the ACCO has responsibility include:

- Receiving and interpreting information. The ACCO receives information from the subscriber network on the receive display (RD) screen. This information is encoded in TACFIRE specific formats, and requires that the ACCO be skilled in interpreting the information and acting quickly.
- Composing and sending messages. The ACCO enters data into the system via special message formats on the compose/edit display (C/ED) screen, then makes proper switch-actions to initiate processing of the data and transmission to designated on-line subscribers.

1.3.2 PLANIT System. The embedded SQT for the TACFIRE system was written in Programming LANguage for Interactive Teaching (PLANIT), a powerful computer-assisted instruction system and courseware authoring language useful for preparing and administering training and evaluation scenarios via computer terminal. Courseware is constructed either interactively or off-line, and the author may execute and/or edit a lesson at any time. PLANIT automatically maintains an audit trail of each student's progress through a course, thus relieving the instructor or scorer of tedious record-keeping responsibilities.

The PLANIT operating system is made machine transportable and has been installed on computer systems supporting FORTRAN, PASCAL, and TACPOL (language used for TACFIRE). For the TACFIRE AN/GYK-12 application, enhancements to the operating system software were implemented to extend PLANIT's capabilities to a specific hardware system. PLANIT's enhanced language for TACFIRE permits the lesson author courseware control of switch indicator lights, and interprets the operation of ACC function keys. Outputs may be directed to either of two screens, to the electronic line printer (ELP), or to the switch panel assembly (SPA). Inputs are processed from the C/ED screen, the alphanumeric keyboard, or the SPA. Under control of PLANIT's enhanced language for TACFIRE, the TACFIRE ACC can operate exactly as it would under control of the operational system software in tactical mode.

2.0 DEVELOPMENT OF COMPUTER-BASED SQT

2.1 Requirements

The development of the HOC offered a unique challenge for complying with regulations that do not explicitly address issues for computer-based SQT testing. Although some steps identified in TRADOC Pam 351-2 were not relevant (for example, training hands-on test scorers and assessing inter-rater reliability), and other additional developmental steps were included, the TRADOC regulation and pamphlet served as adequate guidance for the development process.

2.2 Development Process

The major activities conducted during development of the SQT were:

1. Select tasks for testing.
2. Assign tasks to SQT components.
3. Develop SQT plan.

4. Develop test materials.
   a. Collect task performance data.
   b. Organize hands-on tests.
   c. Program courseware for hands-on tests on commercial computer.
   d. Try out courseware materials on TACFIRE computer.
   e. Revise hands-on tests.
5. Develop support documentation--"Manual for the Administration of the Hands-On Component" (MAHOC), SQT Notice hands-on test worksheets, "Data Supplement," and "Operator's Guide for Preparing SQT Materials."
6. Conduct expert tryout.
7. Submit initial draft.
8. Conduct soldier tryout.

Pertinent characteristics of the finalized TACFIRE embedded SQT HOC are summarized in Table 1. Koch, Pine, and Steinheiser (1980) and Koch, Englert, and Vestewig (1981) provide details of the development process for the SQT.

TABLE 1. CHARACTERISTICS AND CAPABILIITES OF THE TACFIRE EMBEDDED SQT HANDS-ON COMPONENT

Objective
o Assess qualification of artillery control console operator (ACCO), MOS 13C, skill levels 3 and 4.
Conditions of Testing
o Delivered on operational tactical TACFIRE equipment using TACFIRE AN/GYK-12 computer or on TACFIRE Training System (TTS).
o Capability of simultaneous testing of soldiers.
o High fidelity to soldier's manual (FM 6-13C3, FM 6-13C4) tasks.
o Standardized administration at either centralized (Field Artillery School) or field unit installations.
Test Organization
o HOC composed of hands-on tests.
o Automatic presentation of all necessary test instructions and hands-on test conditions, standards, and performance measures.
Test Scoring
o Scoring performed automatically by TACFIRE computer.
o Each hands-on test scored "GO" or "NO-GO" based on number of performance measures passed and satisfaction of time limit standard.
o Qualification on SQT based on criterion percentage of hands-on tests scored "GO."
Test Results Feedback
o Immediate test results and performance feedback.
o Performance feedback specific to items at the performance measure level.
o Individual Soldier's HOC Report providing tabulation of results.
o HOC Record Sheet providing official record of test results.

3.0 VALIDATION OF SQT HANDS-ON COMPONENT

3.1 Validation Plan

The validation phase of the SQT development involved the production of a validation plan and two formal tryouts of the HOC in accordance with the TRADOC guidelines

(TRADOC Pam 351-2). The objective of the validation was to examine the HOC for content validity, accuracy, perceived fairness, and administration feasibility. The strategy for validation of the HOC is reflected in the flow-chart in Figure 2.

Two unique issues were important in assuring an acceptable validation of the computer-based, embedded SQT. The first concerned a necessary distinction between logic errors in the hands-on test courseware and unfair or inaccurate test content per se. Thus, two types of validation were required--one for courseware materials and one for hands-on test technical content. The second issue concerned the requirements for administration (using the resources of a typical unit) and maintainability associated with a computer-based SQT.

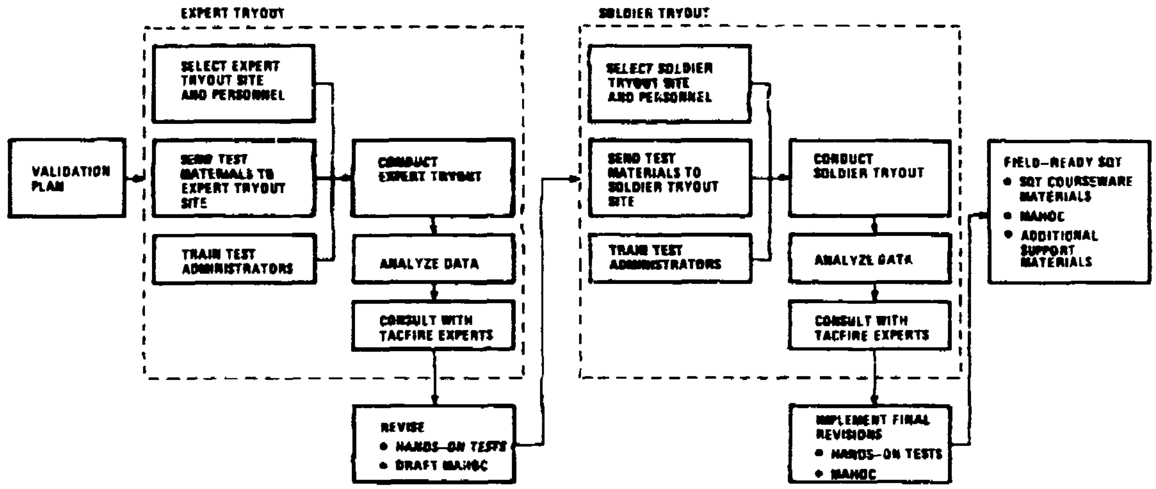


Figure 2. Flow Diagram of SQT HOC Validation Process

3.2 Expert Tryout

3.2.1 Objectives. The objectives of expert tryout of the TACFIRE SQT HOC were to:

- o Provide TACFIRE subject matter experts an opportunity to systematically review each hands-on test for accuracy of the task procedures as defined by the soldier's manual.
- o Examine the draft MAHOC for content, readability, and clarity.
- o Identify and implement any necessary revisions to the hands-on tests, instructions, and draft MAHOC.
- o Determine test administration requirements, including time to set up and administer the hands-on tests.

3.2.2 Procedure. The participants in expert tryout were 14 instructors in the TACFIRE Advanced Training Program at the U.S. Army Field Artillery School, Fort Sill, Oklahoma. Both noncommissioned and commissioned officers were included in the sample. Five experts completed the SQT3 HOC (the HOC for skill level 3), and six experts completed the SQT4 HOC (the HOC for skill level 4). Each of the experts served both as test administrator and examinee. A two-hour workshop was conducted to familiarize the participants with procedures for preparing the computer-based HOC to be administered.

During expert tryout, the participants executed the support procedures routine to prepare the HOC to be administered and then performed each of the hands-on tests in the same manner as an examinee would. Data of various types were collected to meet the stated objectives. Table 2 summarizes the type, purpose, and collection time of these data.

TABLE 2. EXPERT TRYOUT DATA COLLECTION

Type	Purpose	When Collected
Expert Review Form	Examine task domain sampling, hands-on test accuracy, test instructions, and test fairness	End of each hands-on test
Draft MAHOC Review Form	Examine Draft MAHOC	End of workshop to train test administrators
Support Procedures Review Form	Examine test administrator's support procedures	End of test session
General Computer Instructions Review Form	Examine general HOC instructions to the examinee	End of test session
Error Data	Examine hands-on tests	Recorded automatically by computer during test session

3.2.3 Results. Responses to review forms on the draft MAHOC, support procedures, and general instructions identified areas for revision. The section in the MAHOC dealing with computer malfunction recovery was considered inadequate, and certain portions of the general instructions were rated as unclear. Prerequisite data provided on the ELP for some hands-on tests were difficult to read due to poor printer type element quality and caused some undeserved errors. Responses on the expert review forms dealing with technical content of each hands-on test identified problems in these areas: task requirements unclear; only one try allowed for selecting message format switches; optional data entries/blanks not allowed; alternate legal forms of data entry not allowed; answer match lines containing typographical errors.

Analysis of experts' errors on the hands-on tests yielded three identifiable causes of performance measure failures:

- o Operator error--incorrect switch-action or failure to complete hands-on test within appropriate time limit.
- o Courseware logic error--mistakes in the PLANIT courseware that may cause unwarranted performance measure failures.
- o Inappropriate time standard--insufficient time allowed for completing hands-on tests.

Distinguishing operator errors from logic errors required an examination of the printer output for inconsistencies between the expert's response and his detailed test performance feedback.

The remainder of the error analysis focused on the operator errors. Four classifications of operator errors and the average frequency with which they were committed are: a) incorrect data entry--52%; b) incorrect message format selection--23%; c) incorrect switch action--13%; d) excessive time to complete the task--12%.

3.2.4 Revisions. Revisions to the hands-on test materials based on expert tryout data included: expanding the computer malfunction recovery section of the MAHOC; developing a "Data Supplement" containing hardcopy of prerequisite data for all hands-on tests; rewriting portions of general instructions in simpler language; clarifying task requirements; implementing two tries to select correct message format switches; allowing optional data entries/blanks; allowing alternate legal forms of data entries; correcting typographical errors in answer match lines; correcting courseware logic errors; revising inappropriate time limit standards. After modifications to the SQT HOC had been implemented, the hands-on tests were again administered to a subset of the expert tryout participants to verify accuracy of the revisions.

3.3 Soldier Tryout

3.3.1 Objectives. The objectives of soldier tryout of the TACFIRE SQT HOC were to:

- o Determine the HOC's administrative feasibility in a realistic test setting with representative soldiers in the MOS.
- o Continue to examine the hands-on tests for accuracy, appropriateness, and fairness.
- o Perform statistical analyses of hands-on test performance data.
- o Continue to examine the draft MAHOC for clarity and completeness.
- o Implement revisions to the hands-on tests, instructions, and draft MAHOC where necessary.

3.3.2 Procedure. The soldier tryout participants were 22 soldiers in MOS 13C stationed at Fort Sill, Oklahoma. Ten staff sergeants completed SQT3 HOC and ten sergeants first class completed SQT4 HOC. Personnel from the Directorate of Training Development, U.S. Army Field Artillery School served as test administrators.

The type, purpose, and collection time of soldier tryout data are shown in Table 3. The data derived from these sources were analyzed according to the process depicted in Figure 3.

TABLE 3. SOLDIER TRYOUT DATA COLLECTION

Type	Purpose	When Collected
Soldier Review Form	Examine hands-on tests for fairness, accuracy, and clarity of instructions	End of test session
General Computer Instructions Review Form	Examine general HOC instructions to the examinee	End of test session
Error Data	Examine hands-on tests; perform statistical evaluation of HOC	Recorded automatically by computer during test session
Timing Data	Determine test administration requirements; examine performance standards for task completion time	Recorded automatically by computer during test session

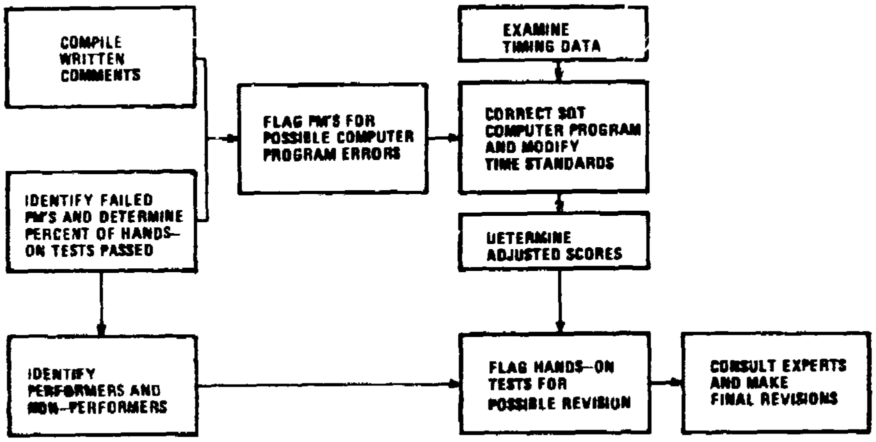


Figure 3. Flow Diagram of Error Analysis and Test Revision Process Following Soldier Tryout

3.3.3 Results. Responses on the general instructions review forms identified two portions of the instructions that some soldiers found unclear--action to take after a discovered error, and amount of help expected from the test site manager. Responses on the soldier review forms evidenced that a large percentage of the soldiers viewed the SQT HOC as a fair and acceptable measure of their ability to perform job duties.

Errors committed on the hands-on tests were again categorized by source--operator error, courseware logic error, and inappropriate time standard. The type and average frequency of errors attributable to the operator are: a) incorrect data entry--51%; b) incorrect message format selection--18%; c) incorrect switch action--21%; d) excessive time to complete the task--10%.

Analysis of the timing data for the representative soldiers provided an empirical basis for examining the adequacy of the hands-on test time limit standards. An average of 83% of soldiers taking SQT3 HOC finished the hands-on tests within given time limit standards; 92% of those taking SQT4 HOC finished within the standards. Three specific criteria were used to judge the adequacy of each hands-on test time standard: a) equals approximately double the average time required by representative soldiers; b) exceeds the upper limit of the range of task completion times; c) equals a value defined by training developers to be mission-related.

3.3.4 Revisions. Revisions to the hands-on test materials based on soldier tryout data included: rewriting portions of general instructions; including SQT HOC administration time guidelines based on soldier tryout in the MAHOC; correcting courseware logic errors; adjusting hands-on test time standards.

4.0 ADMINISTRATION AND MAINTENANCE ISSUES

4.1 Issues in Administering the SQT HOC

TRADOC regulations for SQT development stress the importance of one feature that should characterize an SQT--relative ease of administration (TRADOC Reg 351-2). To meet this criterion, the SQT should be easy to administer by NCOs at the E6 level or higher, be accompanied by comprehensive and simple guidelines for test administrators, and require only the personnel and equipment resources available at a typical unit. These became the objectives for administering the TACFIRE embedded SQT. Issues important for SQT administrative feasibility include qualifications of test administrators, system configuration and status, test security, and disposition of test results.

With the MAHOC as a reference, administering the computer-based SQT is designed to be a turn-key operation for the test administrator. He will be able to perform test set-up procedures explained in the MAHOC, administer the SQT to soldiers, and collect the examinee test results with no prerequisite computer programming experience or outside assistance.

The SQT HOC at either skill level 3 or 4 can be administered on any existing TACFIRE system configuration, that is, battalion or division artillery, at centralized locations or field unit locations. The SQT materials written in PLANIT courseware are transportable via standard TACFIRE input-output medium--magnetic tape cartridge. The load procedures are comparable to those performed by an ACCO for TACFIRE tactical system software. In the event of external threat or other emergency, the TACFIRE system can be restored to operational status immediately by loading tactical system software.

Test security for the computer-based SQT depends on safeguards against compromise by those who are not authorized to administer the SQT and safeguards against unauthorized modifications to the courseware. The PLANIT software provides for both safeguards since log-in identifications used by operators and examinees restrict their access to only those functions in the courseware intended for their use.

For the computer-based SQT, the scoring procedure will differ from the conventional approach. The TACFIRE computer will perform the scoring for each examinee, tabulate test results, and print a complete HOC Record Sheet. This record sheet can be transcribed directly to the SQT marksense form used by the training standards officer to compile test results.

4.2 Maintainability Issues and Procedures

There are four general levels of maintenance for the computer-based, embedded SQT considered important for an overall goal of ease of maintenance. The levels of maintenance are listed in Table 4 ordered from specific to general, together with the primary objective associated with each.

The aspects of TACFIRE equipment maintenance relevant to actual system operation also affect SQT operation. In the event of equipment malfunctions, the testing is halted but no performance data are lost from core memory. When the system is restored, PLANIT automatically resumes HOC testing at the point where each examinee was stopped.

Annual revisions to an SQT mandated by TRADOC are accommodated by a large base of validated hands-on tests. The SQT developers need only select a different subset of hands-on tests each year to optimally represent extant job duties. This selection requires no modifications to the courseware.

TABLE 4. MAINTENANCE OBJECTIVES

Level of Maintenance	Primary Objective
1. Maintenance for TACFIRE equipment on which the SQT HOC is administered.	Ensure test reliability by minimizing equipment failure and down time.
2. Annual revisions to SQT contents in accordance with TRADOC regulations.	Provide flexibility to use developed and validated hands-on tests for more than one test period.
3. Modifications and additions to the courseware comprising hands-on tests.	Provide capability to implement modifications and additions to the courseware interactively, easily, and rapidly.
4. Enhancements and troubleshooting support for the software system.	Provide quality assurance for system software.

New interface requirements and operating system software releases will drive the training and testing requirements for TACFIRE and any evolving, complex system. Supplementing the TACFIRE SQT with new or revised test materials is an important capability to accommodate extensions in the scope of the SQT. A technician with computer programming experience has the qualifications to edit existing courseware and develop new courseware for the TACFIRE embedded SQT. The more common types of courseware modifications that will be required are quite simple and can be accomplished interactively from the ACC terminal at any time.

At the most general level of maintenance for the computer-based SQT is support of any enhancements for the PLANIT operating system software. The evidence for maintainability and supportability of the PLANIT software is derived from many trouble-free installations of PLANIT nationwide, spanning a wide variety of applications.

5.0 CONCLUSIONS

Conclusions warranted by the results of the computer-based SQT development and validation may be relevant both as an overview of the present work and as research and development issues for other SQTs designed to be delivered on operational tactical data systems. First, the TACFIRE embedded SQT appears to be enthusiastically accepted by the user community. The extensive involvement of subject matter experts and TRADOC personnel during the development, validation, and review process, together with the developer's responsiveness to TRADOC regulations and user input, are thought to be important contributors to this acceptance.

The TACFIRE application of computer-based, embedded HOC testing is feasible and effective in fulfilling the intent of the Army SQT program. Performance of hands-on tests on the TACFIRE equipment closely resembles the performance of actual job duties. The TACFIRE computer accomplishes objective and accurate scoring of hands-on tests. The SQT is also feasible from the standpoint of administration and maintenance, requiring no programming experience to administer and no other resources or equipment than those available at the typical field unit. The PLANIT courseware materials are structured in a modular format and are easy to modify and supplement.

Compliance with TRADOC regulations for SQT policy and procedures by the computer-based SQT does not differ significantly from conventional SQTs. The validation plan, expert tryout, soldier tryout, data analysis, and test content revisions were conducted in accordance with regulations. The MAHOC closely follows the TRADOC recommended format. Test results are computer-generated in a form compatible with TRADOC SQT scoring procedures.

PLANIT has been a successful vehicle for development of the TACFIRE embedded SQT. PLANIT capabilities and advantages include: extensive record keeping of student performance; analysis of incorrect answers to provide detailed feedback; transportability of courseware; ease of courseware modification. The success of PLANIT in the implementation of the TACFIRE embedded SQT recommends it as a candidate for other developments on computer-based tactical data systems.

REFERENCES

Guidelines for Development of Skill Qualification Tests. TRADOC Pamphlet No. 351-2 (Draft), Department of the Army, Headquarters, U.S. Army Training and Doctrine Command, Fort Monroe, VA, 1980.

Koch, C.G., Englert, J.A., & Vestewig, R.E. Development and Validation of a Computer-Based Skill Qualification Test--Final Report. Report No. 81SRC48, Honeywell Systems and Research Center, Minneapolis, MN, September 1981.

Koch, C.G., Pine, S.M., & Kingsbury, G.G. Development of a Computer-Based Skill Qualification Test--Final Report. Report 80SRC52, Honeywell Systems and Research Center, Minneapolis, MN, July 1980.

Koch, C.G., Pine, S.M., & Steinheiser, F.H. "Development of a Computer-Based Skill Qualification Test." Proceedings of the Military Testing Association 22nd Annual Conference, Toronto, Ontario, 27-31 October 1980.

Skill Qualification Test (SQT) Policy and Procedures. TRADOC Regulation No. 351-2, Department of the Army, Headquarters, U.S. Army Training and Doctrine Command, Fort Monroe, VA, 21 April 1980.

Soldier's Manual, TACFIRE Operations Specialist, MOS 13C, Skill Level 3. Field Manual No. FM 6-13C3, Headquarters, Department of the Army, December 1980.

Soldier's Manual, TACFIRE Operations Specialist, MOS 13C, Skill Level 4. Field Manual No. FM 6-13C4, Headquarters, Department of the Army, December 1980.

Lessard, Pierre, CPT, Canadian Forces Personnel Applied Research Unit, Willowdale, Ontario. (Thurs. A.M.)

A Simulated Aircraft Landing Test as a Pilot Selection Test

Pilot selection in the Canadian Armed Forces has been a continuing concern especially since attrition rates for trainees has been as high as fifty percent on a few courses. In the reported study, the utility of a custom-developed, microprocessor-driven aircraft landing test (ALT), was examined in terms of its added value to the current pilot selection battery. A CROMENCO II microcomputer was programmed to simulate the landing of a light aircraft where visual stimulation was presented on a CRT and the candidates used a "joystick" and throttle to perform three landing tests. Several dependant measures were automatically recorded every 500 m sec. This ALT was a further development of earlier research conducted on another system. (Fowler, 1981).

One hundred and fifty male candidates applying for military flying training were tested on the ALT as well as on the current test battery which consists of a psychomotor test in an aviation tester, pencil-and-paper tests which tape verbal and quantitative aptitudes and a memory test. The candidates also completed a measure of cognitive style, the Group Embedded Figures Test, and selected scales from Jackson's Personality Research Form.

It was found that performance on the ALT was independent of performance on the current test battery as well as performance on the Croup Embedded Figures Test and the selected scales from Jackson's Personality Research Form.

Future research into the use of the ALT as a selection device will take place once trainees have completed the primary flying training course.

AN INVESTIGATION INTO THE USE OF AN AIRCRAFT LANDING TEST (ALT) IN AIRCREW SELECTION

Captain P.B. Lessard
Canadian Forces Personnel Applied Research Unit
Suite 600, 4900 Yonge Street
Willowdale, Ontario M2N 6B7

INTRODUCTION

Background

The Canadian Armed Forces (CAF) has a continuing need to improve their Aircrew Selection Battery (ASB) to ensure its predictive efficiency. The importance of psychomotor skills, (i.e., the ability to co-ordinate, manipulate and/or repeat precise body or limb movements) and to respond to visual stimuli, has long been recognized as critical for success as a pilot. The United States Air Force (USAF) in a study of perceptual motor and cognitive performance (Imhoff & Levine, 1981), found the predictive validity of the various perceptual-motor and psychomotor tests used in pilot selection to be the most significant.

A study of student pilot attrition (King & Eddowes, 1976) in the USAF Undergraduate Pilot Training (UPT) program found that failed students typically listed pre-solo landing, loss of confidence and final turn-approach-flare as areas which lead to their subsequent elimination from pilot training. During personal interviews with Royal Military College cadets who had no previous flying experience, Capt. G.A. VanDyke (Memorandum, March 1981) of the CAF reported that 10 of the 24 failed cadets noted that landings were by far the most difficult part of the course. These results indicate that perhaps an aircraft landing test might help identify potential failures at the selection stage.

Also, the CAF is in the rather unusual position of testing and selecting persons for pilot training from different cultural and language groups within the country. The use of performance tasks rather than pencil-and-paper tests could help eliminate cultural biases which are often found in translated tests.

Hypothesis

The hypothesis is that there is a significant positive correlation between performance on the ALT and performance on the ASB, particularly with the Visual General Aviation Tester (VGAT).

METHOD

Subjects

The ALT was administered to a sample of 107 male Anglophones between the ages of 17 and 26, all applying for flying training in the Canadian Armed

Forces. None of the applicants had previous flying experience. The education level of the applicants varied from at least a Junior Matriculation (high school graduates) to Direct Entry Officers (DEOs), those with a college or university education. All applicants had been prescreened using the Canadian Forces General Classification (CFGC) test and had undergone an initial interview to assess officer potential. All subjects, whether or not selected for pilot training, were used in the analyses of the ALT.

Apparatus

The ALT, driven by a Cromemco microcomputer was used in conjunction with a cathode ray tube (CRT) display to simulate the approach and landing of a light aircraft. The throttle was positioned in front of the control column to accommodate both left- and right-handed persons. An adjustable chair was used to ensure all subjects were looking directly at the CRT. Approach, stall and structural damage speeds were noted on the left-hand side of the console as reminders to the subject.

The simulation was that of a cross (the aircraft), which was to be flown through an approach to a safe landing using as instruments, an horizon bar, airspeed indicator and altimeter which were controlled by throttle and control column movements. The approach commenced at 900 feet altitude, 2700 feet from the threshold of the runway. The approach was terminated when a safe landing, a stall (air speed less than 50 mph), a crash (airspeed in excess of 75 mph or less than 65 mph on landing) or structural damage occurred (speeds in excess of 150 mph). All parameters were recorded on hard copy using a Texas Instruments printer. Figure I is a block diagram depicting the CRT and accessories which are included in the design of the ALT.

Procedure

Subjects, who are processed through the Aircrew Selection Centre (ASC), are selected for pilot training based on the results from a battery of pencil-and-paper tests and a psychomotor test, the VGAT. The ALT was administered to the subjects as part of the ASB without advising them it was an experimental test to ensure they made every effort to do as well as possible. The results obtained on the ASB, or the ALT, were not released to the subjects.

The administration of the ALT, which took approximately 30 minutes consisted of a briefing given to all applicants and a testing period. The briefing consisted of an outline of all instruments and controls, plus a demonstration of all controls, their function, airspeeds to monitor symbology and warnings that could appear on the screen. It should be noted that the ALT had no rudder control, and all turns were made by banking. The instructor also demonstrated a successful landing. Subjects were tested individually.

At the conclusion of the demonstration the subject had the opportunity to practise co-ordinating the controls and flying to a targeting circle. Once this familiarization exercise was complete the candidate attempted three landings which were scored, keyed and stored on disc.

RESULTS

The ALT scores were recorded as successful versus failed (stalled or crashed) a dichotomous variable. In addition total time for the approach, time outside the approach cone, time outside the approach speed, and distance from the runway threshold when a successful landing, crash or stall occurred were also scored.

There was no correlation between the number of successful landings (mean = 1.44, standard deviation = 1.09) and scores from the Aircrew Selection Centre battery pencil-and-paper tests. However, a significant relationship was found between the Visual General Aviation Test (VGAT) and the ALT average landing accounting for approximately four percent of the vaiiance.

Table 2, a t-test using extreme groups (those with zero successful landings and those with all three landings successful) showed that there is a significant difference between the two groups on one test, serial addition. The remaining seven tests showed no significant difference between the two extreme groups.

DISCUSSION

This preliminary study failed to show a significant positive correlation between scores on the ALT and scores on the ASC battery, specifically VGAT. Since both the ALT and VGAT are psychomotor tests it was felt there should be some relationship between the two. This study failed to support the hypothesis. One of the major problems encountered in using the ALT as a testing device is its lack of fidelity or the precision with which a landing could be reproduced on the CRT. Furthermore, visual feedback to the subject (i.e., airspeed, altitude, attitude) and kinesthetic feedback (control column and throttle movement) are far from the quality necessary to duplicate landing a light aircraft. In general, the graphics package and mechanical controls require considerable work prior to conducting another study. Specifically, the legibility of instructions and feedback statements detract from the tests reliability as a selection device. Before dismissing the ALT out-of-hand however, flying training data from the Canadian Forces Flight Training School will be obtained and the ALT scores correlated with them. Therefore, a further investigation will be carried out in early 1982, when sufficient numbers of candidates have completed basic flying training to further examine the relationship between flying training and the ALT.

References

Imhoff, D.L. & Levine, J.M. Perceptual-motor and cognitive performance task battery for pilot selection. AFHRL-TR-80-27, 1981.

King, N.W. & Eddowes, E.E. Similarities and Differences Among Superior, Marginal and Eliminated Undergraduate Pilot Training Students. AFHRL-TR-76-12.

Vandyke, G.A. Report on visit to Royal Military College. Personal Communication, March 1981.

TABLE 1

PEARSON CORRELATIONS BETWEEN ALT AVERAGE LANDING
AND ASC TEST SCORES FOR THE MALE ANGLOPHONES

ASC TESTS	PEARSON CORRELATIONS
Numerical Ability (AB)	.02 (n=102)
Verbal Aptitude (TV)	.10 (n=102)
Arithmetic Reasoning (AN)	.09 (n=102)
Reading Comprehension (TR)	.05 (n=102)
Math Reasoning (AR)	.08 (n=102)
Instrument Interpretation (WC)	.16 (n=102)
Reading Tables (WT)	.14 (n=102)
Serial Addition (AS)	.16 (n=102)
VGAT Grand Total Score	.20* (n=101)
General Classification (GC)	.05 (n= 59)

* $p < .05$

TABLE 2

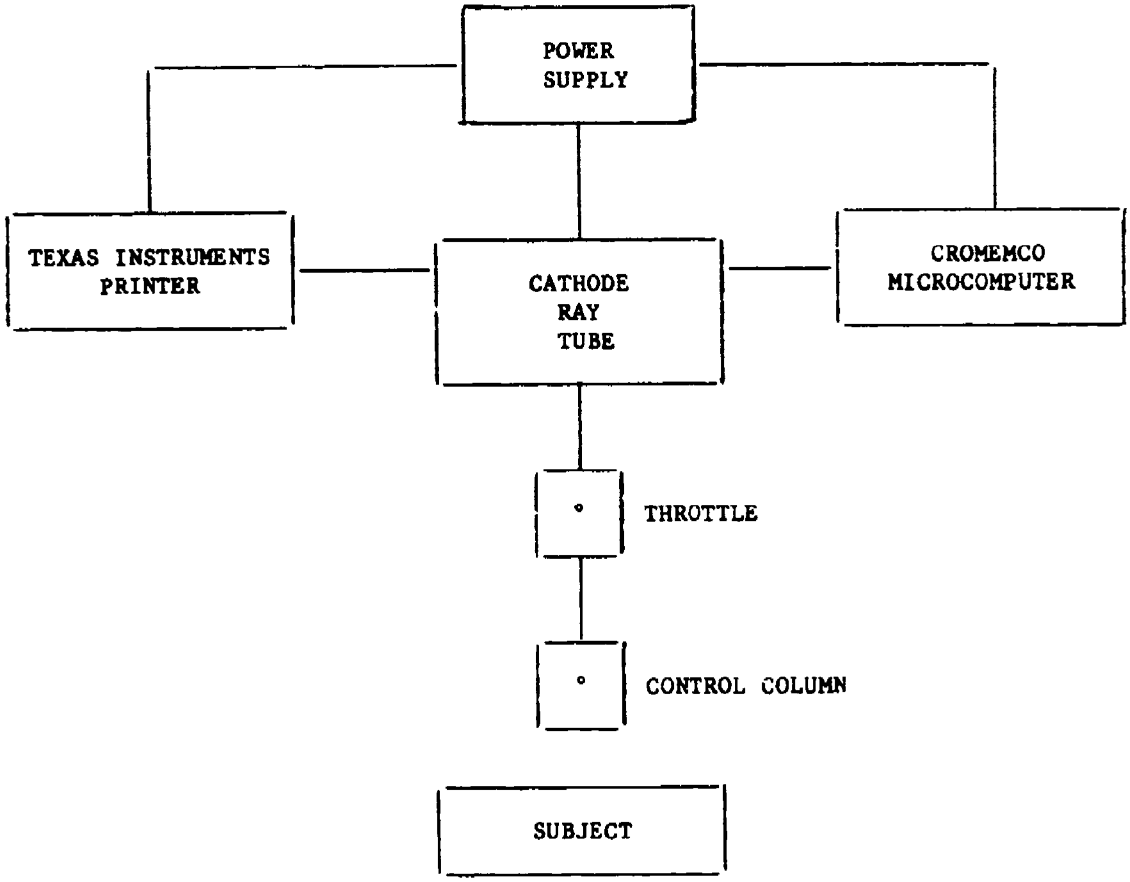
T-TEST ON SERIAL ADDITION SCORES
FOR MALE ALGLOPHONE GROUPS BASED ON ALT PERFORMANCE

VARIABLE	NUMBER OF CASES	MEAN	STANDARD DEVIATION	POOLED VARIANCE ESTIMATE		
				T VALUE	DEG OF FREEDOM	TWO-TAIL PROBABILITY
SERIAL ADDITION						
GROUP 1	28	33.79	9.56	-2.31	46	$p < .05$
GROUP 2	20	39.95	8.44			

Group 1 - extreme group which did not successfully land the ALT at all

Group 2 - extreme group which successfully landed the ALT on all three attempts

FIGURE 1


POWER SUPPLY
TEXAS INSTRUMENTS PRINTER
CATHODE RAY TUBE
CROMEMCO MICROCOMPUTER
THROTTLE
CONTROL COLUMN
SUBJECT

FLYING SELECTION IN THE ROYAL AIR FORCE*

I. G. LIDDERDALE
RAF Exchange Psychologist, Operations Training Division,
USAF Human Resources Laboratory, Williams AFB, Arizona.
&
C. C. ELSHAW
Research Branch, HQ RAF Support Command
RAF Brampton, England.

SUMMARY

The paper describes the results of a study designed to evaluate the utility of a flight selection test for Royal Air Force pilot candidates. 53 pilot applicants acted as subjects in a controlled trial, based on a 14 hour light aircraft course. Tests were marked in the air at 9 and 14 hours by independent examiners. When all of the students had completed the RAF Basic Flying Training course one year later, their tests results were compared with training outcomes. The results indicated a very high relationship between flying test marks and probability of success in later training. The marks awarded by the examiners were more predictive than the assessments of the flying instructors. Based on the results of the trial, the RAF established a Flying Selection Squadron and the paper presents some data from follow-up studies which were carried out as a validation of the selection procedure.

INTRODUCTION

Early in 1942, the Royal Air Force (RAF) introduced a change in its method of selecting pilot cadets for pilot training. The modified plan involved systematic assessment of flying performance after a limited amount of standardised dual flying instruction.

The introduction of this more analytical and objective approach was a radical departure from the previous selection methods which were based largely on a general subjective impression of a candidate's suitability.

This method of selecting potential pilots from aircrew candidates on the results of a standard 'flight test' procedure came to be known in the RAF as Grading.

The introduction of Grading as a method of pilot selection in the RAF was, in part, due to the transfer early in 1942 of the bulk of pilot training overseas. The war situation at that time made it imperative that only cadets who gave very high promise of successful training should occupy shipping space, (Air Ministry Report 1945).

* *Note:*
Any views expressed are those of the authors and do not necessarily represent those of the AOC-in-C RAF Support Command or the Chief Scientist (RAF).

As a result of the introduction of Grading, attrition rates at all stages of training were reduced significantly, particularly at the Elementary Flying Training School (EFTS). The pilot aptitude ratings from the Grading tests showed a high relationship not only with probability of failure but also with the more discriminating criterion of proficiency assessments.

The post-War decline in the requirement for aircrew led to a shift in the selection requirements of the RAF. The relatively expensive procedure of flight Grading, which had been justified by the Wartime requirement, was no longer deemed suitable to the needs of a contracting Service. For many years, aircrew for the RAF were selected using a traditional pilot aptitude test battery.

Grading was re-introduced briefly in the early '50s, with civilian pilots acting as instructors and testing officers under contract to the Government. However, poor standardisation and commitment on the part of the staff of Grading schools resulted in unreliable testing practices and the scheme was abandoned in favour of the less expensive and more reliable aptitude battery, (Ministry of Defence Publication, 1969).

THE PRIMARY FLYING GRADING TRIAL

In 1973, the RAF introduced an all-jet flying training system; as a result, the Primary Flying School (PFS), which had previously provided initial flying training using a Chipmunk aircraft, was closed. This change in the training sytsem came at a time of great concern over the rising attrition rates in the Basic Flying Training Schools (BFTSs) and an equal concern over reports of falling predictive validity of the pilot aptitude tests. The net effect of these concerns was that an experimental flying unit was formed with the objective of testing the concept of reintroducing some form of flying selection.

The Primary Flying Grading Squadron (PFGS), as the unit was called, came into being at RAF Church Fenton in Yorkshire on 6 May 1974. The squadron was formed from the disbanded PFS and was able to capitalise on the availability of experienced flying instructors and support services of the old flying school.The unit consisted of 6 Qualified Flying Instructors (QFIs), 4 as instructors and 2 as examiners, and 5 Chipmunk aircraft.

The aim of the PFGS was to test the feasibility of using a short, light aircraft, flying course as a predictor of later success in RAF flying training.

The PFGS syllabus was developed largely along the lines of the Canadian Armed Forces Primary Flying Grading Squadron, 3 CFFTS, at Portage La Prairie. Inputs were also made from the Federdal German Air Force flying selection squadron at Furstenfeldbruk and the RAF Central Flying School.

The ground school programme of the course was designed to give the students what was considered to be the minimum knowledge of airmanship and aerodynamics necessary to complete the flying syllabus. In addition, each flying exercise had an associated reading assignment which the student had to complete before attending the relevant pre-flight briefing.

The flying syllabus laid great emphasis on 'attitude' flying, high workload and maximum handling of the aircraft by the student from an early stage. To this end, traditional basic handling exercises were compressed and only taught in the way in which they were to be applied in the tests. The course was 15 flying hours, including a flexible allowance of one hour of non-instructional flying hour per student.

Instructor standardisation was rigorously pursued before and throughout the trial using mutual check rides. Examiner standardisation was an even more difficult task as the pilots had to familiarise themselves with a complex test booklet which was required to be marked, in the air, during test flights. Several weeks of training were needed before the examiners were able to reach acceptable levels of inter-rater reliability and checks were made continuously throughout the trial period.

Social interaction between the examiners and the students was kept to a minimum prior to testing and the examiners were not permitted to see student training records. In this way, it was hoped that the tests would not be biased by personal likes and dislikes or knowledge of a students successes or failures during the instructional phase of the course. The tests were designed to be as objective as possible and for much of the flight the examiner acted as a human flight data recorder, noting airspeed, bank and heading deviations.

The trial took almost a year to complete and, during this time, 53 students were tested. Two tests were administered, one at 9 hours and one at 14 hours. All results were passed directly to the Research Branch at HQTC as they became available and they were kept secret until the last of the trial students had completed the RAF 100 hour Basic Flying Training School (BFTS) course. No students were failed as a result of taking part in the trial and their marks were never passed to their flying units nor were they used to make any decisions regarding their careers.

Figure 1. shows the predictive accuracy of the assessments. The white columns show the biserial correlations of the cumulative subjective marks which were awarded by the QFIs. The shaded columns show the biserial correlations of the test marks alone.

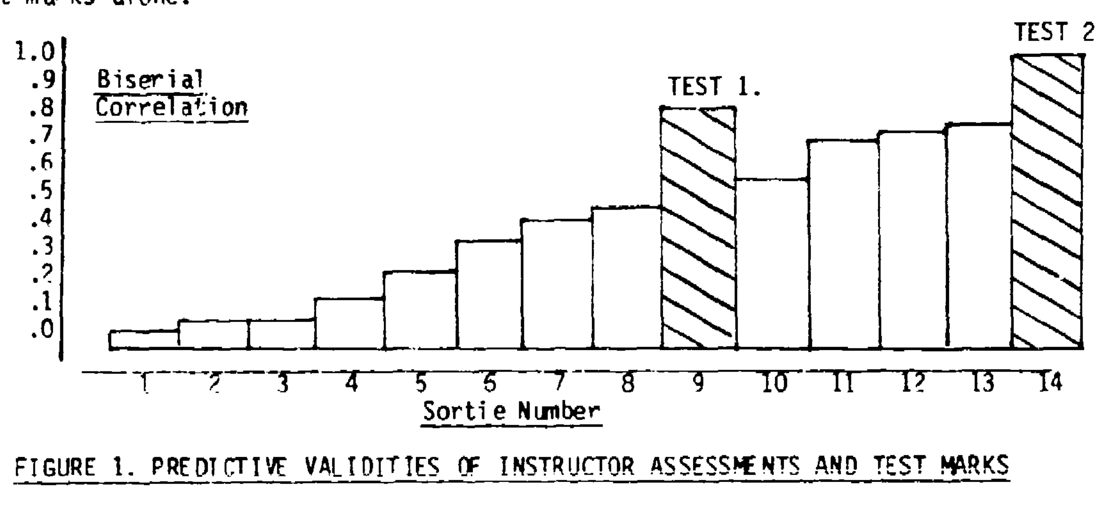


FIGURE 1. PREDICTIVE VALIDITIES OF INSTRUCTOR ASSESSMENTS AND TEST MARKS

It can be seen that, as the course progressed from sortie 1 to sortie 14, the predictive accuracy of the assessments increased, although there was some sign that they were levelling out. The tests, however, show a significantly higher level of prediction at the same point in the course than the cumulative assessments of the instructors after many hours of close contact with the students.

At the end of each course, the instructors and examiners compared notes for the first time and decidee on an overall Grade for each student on a scale from 1 to 10; it was nominally decided that level 5 should represent the dividing line between predicted success and failure.

The Grade was based on a consensus opinion about the likelihood of success in BFTS. Although the Grades were largely influenced by the test marks, there were some occassions where borderline cases were decided on the basis of the instructors' knowledge about the personal suitability of the student for training within the RAF or other overriding reasons why the test marks should be modified, such as airsickness or poor continuity.

Figure 2. below shows the distribution of the final Grades for the 53 students. The circles coloured black represent those who failed subsequent Basic Flying Training. It can be seen that the overall Grades gave a very accurate assessment; 18 students had been predicted as failures and, by the end of the following year 16 had indeed failed. 31 out of the 35 predicted successes subsequently passed Basic Flying Training.

	10	
	9	
Predicted as successes	8	0000000
	7	000000000000
	6	●●●●000000000000
	5	●●●●●●●●●00
	4	●●●●
Predicted as failures	3	●●●
	2	
	1	

0 = Students who passed training.
● = Students who failed training.

FIGURE 2. DISTRIBUTION OF GRADES SHOWING THOSE STUDENTS WHO SUBSEQUENTLY FAILED BASIC FLYING TRAINING SCHOOL

Part of the rationale for introducing a flying selection trial at this time was that the pilot aptitude tests, then in use at OASC, were unable to make sufficiently accurate discriminations between the pilot applicants to ensure low attrition rates at the RAF Basic Flying Training Schools. The validation of the flying selection tests rested therefore, in part, on the comparison of the predictive accuracy of the two measures. In particular, it was hoped that the flying tests would complement the existing selection methods by showing a low correlation with the pilot aptitude battery.

Figure 3. below shows the distribution of pilot aptitude scores for the 53 trial students. The low predictive accuracy of the pilot aptitude scores at this end of the scale is evident. Although the pilot aptitude tests had almost certainly eliminated many totally unsuitable candidates who scored at the lower end of the complete aptitude scale, it was clear that further discrimination was not possible within the remaining applicants without further testing. The flying selection tests proved that this finer level of discrimination was possible.

```
                                    161-170 |  0
                                    151-160 |  ●0000
+------------------------------+    141-150 |  ●●●●●000000
| 0 - Students who passed BFTS |    131-140 |  ●●●●●●●●0000000000
| ● - Students who failed BFTS |    121-130 |  ●●●●000000
+------------------------------+    111-120 |  ●00000
                                    101-110 |  ●
```

FIGURE 3. RELATIONSHIP BETWEEN PILOT APTITUDE SCORES AND SUCCESS IN BASIC FLYING TRAINING

Following the publication of the results of the Primary Flying Grading Trial, extensive cost-benefit analyses were carried out to determine the effect of introducing flying selection into the RAF. A computer-based manpower and cost model was devised specifically for this analysis. The model was able to provide accurate cost and resources figures which were used to assess the impact of adding an extra stage in the selection and training system of the RAF.

The model was also used for to assess alternative solutions to the pilot selection and classification problem, such as simulator selection or improvements in the pilot aptitude battery.The analyis showed that a flying selection system would be highly beneficial in terms of both costs and manpower, particularly when it is used to supplement existing testing and training procedures.

It had been known for some time that students who entered the RAF with more than 30 hours of previous flying experience had a higher chance of success than 'ab initio' pupils. The Flying Grading Squadron had little to say about such candidates, and in addition was not designed for them. It was therefore decided to eliminate the need to test such candidates further after initial pilot screening at the Officer and Aircrew Selection Centre.

Students who enter the RAF through the University Cadetship sytem receive flying training on a Bulldog aircraft during their college years. These students were also considered unsuitable candidates for additional flying selction.

Having established how flying selction could be most profitably employed within the constraints of the RAF training and selection system a decision was made to establish Flying Selection as a filter for candidates with less than 30 hours of previous flying experience who were not University Cadets.

THE RAF FLYING SELECTION SQUADRON.

The RAF Flying Selection Squadron (FSS) was established in September 1979 at RAF Swynderby, using Chipmunk aircraft which had been in storage for such an eventuality. The procedures of the FSS are largely those used during the trial with some modification to the test content as a result of detailed item analysis.

The course is scheduled for 6 weeks and comprises 10 or 11 candidates. The 14 hour syllabus has test points at the 9th and 14th exercise, each marked by a different examiner; there is an instructor change after the first test. The flying instructors use subjective assessments to describe the students' performance after each exercise and at the end of each phase of the course, the instructor makes a prediction of the candidates' probability of success in later flying training.

The tests are marked in the air by independent examiners. The various items in the test are weighted and combined to produce an overall score. The examiners also give a subjective rating of the students' performance.

The decision to either pass or fail a student is taken as a result of a Disposal Board meeting which is convened aa soon as possible after the completion of a course. The Board is Chaired by the Group Captain responsible for RAF Flying Training at the Command Headquarters.

The executive members of the Board are the Officer Commanding the FSS and the 2 examiners. Advisory members include all of the flying instructors and a Command Research Branch psychologist. The Board has, in attendence, a representitive from the Command Headquarters Personnel staff to advise on alternative carreer placement for the failed students.

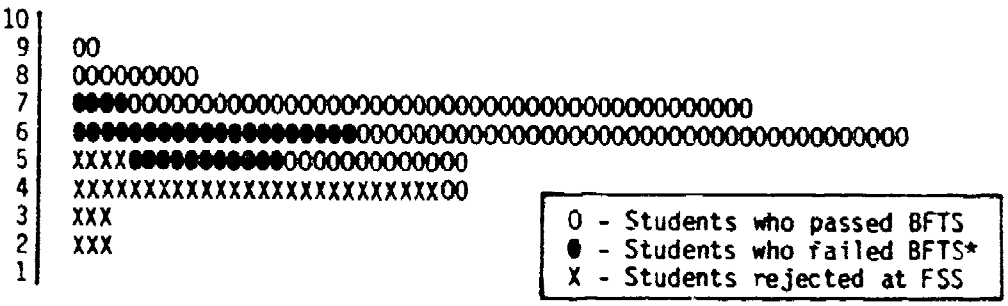
The Board notes the subjective assessments given throughout the course and takes into account any special considerations such as sickness, personal problems, poor flying continuity or adverse weather during a test sortie. Initial subjective Grades are recorded and then compared with the objective rank order; at this time any anomalies between the two are resolved. In this way, a mixture of objective and subjective assessments produces a final FSS Grade in which neither element overwhelmingly dominates the other.

At this stage any anomalies between the test marks and the subjective opinions are resolved. The Chairman acts as the final arbiter for all borderline cases.

FSS has now been running for 2 years, during which time 352 candidates, comprising 34 courses, have been processed. 88 of these candidates did not go on to Basic Flying Training; this figure includes 8 students who withdrew from the system at their own request. Thus, the overall FSS rejection rate is approximately 25%. The present policy sets the minimum acceptable Grade at 6 although the Chairman still retains the authority to pass students with a Grade of 5 provided that he is satisfied that the training risk is justifiable.

To date, some 19 FSS courses have completed BFT and this sample of 153 students is sufficiently large to indicate that FSS is performing well.

Distribution of FSS Grades. Figure 4. below compares the FSS Board Grades with overall BFT results. The overall level of prediction, as estimated from these data is very much in accordance with expectations based on the trial results. Clearly, one would expect some lowering of prediction with a real world system due to the problems in maintaining high reliability of assessments and drift in the marking standards, but the results so far are most encouraging.



*failure for airwork only.

FIGURE 4. THE RELATIONSHIP BETWEEN FLYING SELECTION GRADES AND RESULTS AT BASIC FLYING SCHOOL

Effects of FSS on BFTS wastage rates. Table 1. below shows the Basic Flying Training School wastage rate before and after the introduction of the Flying Selection test. The results of this initial validation are most encouraging, and reflect closely the predictions of the computer model of the selection and training system referrred to earlier.

Table 1.	Basic Flying Training School suspension rate.*
Before the introduction of Flying Selection	23%
After the introduction of Flying Selection	16.1%

* Note: these rates only apply to none University cadets with less than 30 hours flying experience before joining the Service. The overall BFTS suspension rate is different from these figures.

Relationship between FSS Grades and Pilot Aptitude Assessments. All candidates who enter the FSS have been previously screened using the Pilot Aptitude battery at the Officer and Aircrew Selection Centre at RAF Biggin Hill. The Pilot Aptitude index is a composite score of paper and pencil, complex co-ordination and other tests. Figure 5. below shows the relationship between FSS Grades and the Pilot Aptitude index of the candidates.

150	X0000000000000000
141-150	XXXX●●●●0000000000000000000
131-140	XXXXXX●●00000000
121-140	XX●●●0000000000000000
111-120	XXXXX●●●●●●●●000000000000000000000
101-110	XXXXXXXXXX●●●●●●●●●●●00000000000000
100	XXXXXXXX●●●●●●●0000000

0 - Students who passed BFTS
● - Students who failed BFTS*
X - Students rejected at FSS

* failure for airwork only.

FIGURE 5. THE RELATIONSHIP BETWEEN PILOT APTITUDE SCORES AND SUCCES IN BASIC FLYING TRAINING. (Courses 1 - 19).

It can been seen that more FSS rejections and BFTS airwork failures now occur amongst those students with aptitude scores of 120 and below. This result is in line with the improved level of prediction from the current pilot aptitude scores as a result extensive test development. However, in spite of this improvement in the aptitude testing battery, FSS continues to have an additional discriminating effect as shown by Figure 6 below which illustrates that the Flying Selection testing procedure is identifying potential BFTS failures across the full range of aptitude scores. BFTS failures have not been included in this Figure as many of the students are still under training.

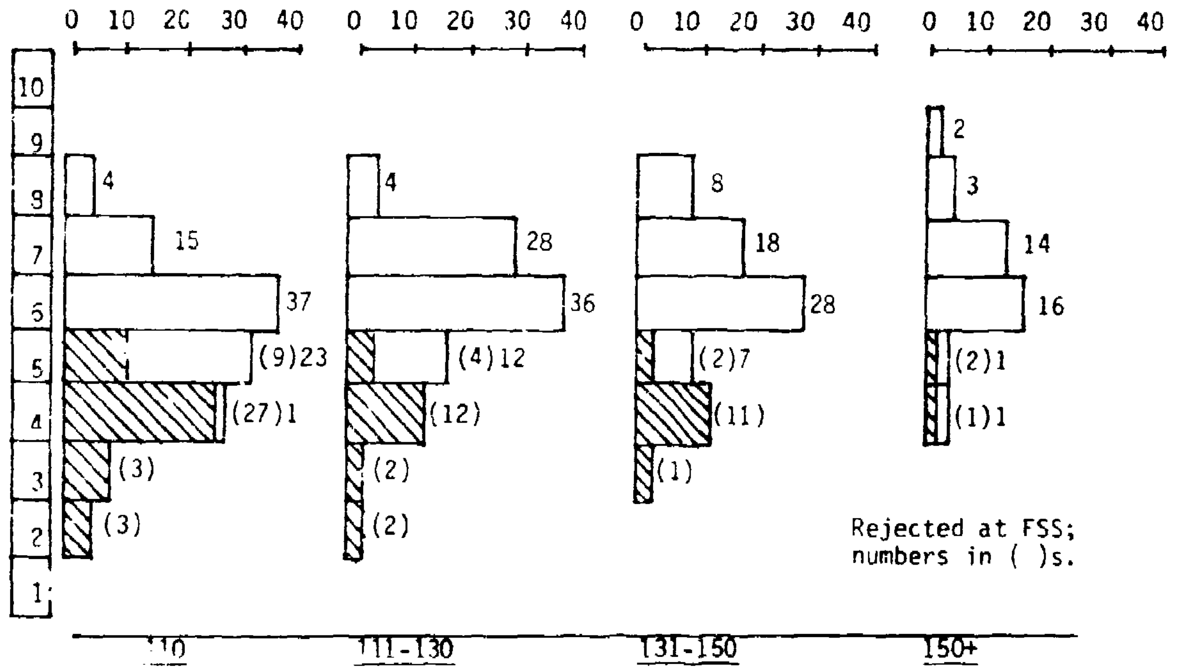


FIGURE 6. RELATIONSHIP BETWEEN PILOT APTITUDE SCORES AND FLYING SELECION SQUADRON REJECTIONS. (Courses 1 - 35).

Student retention and manpower implications. The present RAF Officer and Aircrew Selection Center tests make an initial division of aircrew applicants into pilots and navigators as a result of their performance on specific selection criteria. An alternative option would be to make this placement decision after Flying Selection and this option is currently under investigation.

Effects on navigator targets, standards and morale will, of course, have to be taken into account, but the high transfer rate of rejected FSS candidates gives cause for some optimism. Out of 88 FSS rejections so far, only 6 have left the Service; a remarkable 93% retention rate. The majority, 60, were transferred to navigator training; 13 went to ground branches and disposal action for the remaining 9 is still being considered.

It would appear that, having obtained their commissions and put some useful Service experience behind them, rejected FSS candidates are motivated to remain in the Service, even if they had originally expressed the intention of being a pilot or nothing. This tends to support the view that Flying Selection should be established after, rather than before Initial Officer Training.

CONCLUDING REMARKS

Since the end of the second World War, an enormous amount of time and effort has been expended in the search for improved pilot selection tests. However, despite this considerable and continuing research investment, it has been said that the major advances in testing in the last 25 years have been in the area of statistical methodology rather than in test content (North and Griffen, 1977).

It is also worth noting that the development of tests of high predictive validity is not, in itself, sufficient for them to be included in a pilot test battery. As an example, in 1951 the US Air Force gave up the use of psychomotor tests in selection in spite of their unique contribution to selection validity (Cronbach 1970). Since that time many other selection measures which have demonstrated high predictive validity have been excluded from pilot selection batteries because of administrative, reliability and quality control problems.

While there was little doubt that selection tests based on light aircarft performance were able to produce the necessary levels of predictive validity, the development of a practicable and reliable test format was a considerable challenge.

In spite of the complexity of the flying environment, the test methods used in the Primary Flying Grading Trial in 1974 produced very high rater reliability. This finding is in line with similar studies, such as the development of the 'Illinois Private Pilot Flight Performance Scale' (Povenmire, Alvares and Damos 1970) in which high observer-observer reliability was observed in a student grading test situation.

The justification for introducing Flying Selection into the RAF was, however, based on a very detailed analysis of the costs, benefits and manpower implications of multi-stage selection systems. Without a detailed front-end analysis of this kind the full benefits of Flying Selection may not have been apparent. Flying Selection is, after all, an expensive form of testing, and it is not suitable or appropriate for all candidates. However, it is also highly predictive and, so far, reliable and administratively convenient.

Many questions remain. Why is it, for example, that students who have flown before joining the Service do so much better in training? It is clear that they have skills which will be of benefit during their military flying training but the research evidence tends to suggest that the important difference may lie in their motivation towards a flying career.

Flying Selection provides an opportunity for the Air Force to take a closer look at the abilities of candidates before entering them into a very costly training system. It also allows the candidates with no experience of flying to assess their own abilities and motivation within a supportive environment.

After two years of using Flying Selection, the RAF has proved that the use of standardised flying tests can add a level of discrimination to the pilot selection process which is not obtainable in any other way. Maintenance of reliability will, undoubtedly, be the greatest challenge in the years to come but, based on our current experience, the rewards will justify the effort.

REFERENCES

Air Ministry Pamphlet 190: Grading - A report on pilot selection in the Royal Air Force, 1945.

Cronbach, L. J. Essentials of Psychological Testing, 3rd Edition, Harper and Row, Inc., New York, 1970, pp. 97.

Ministry of Defence Publication: History of Officer Selection in the Royal Air Force, 1958-69.

North R. A. & Griffin, G. R. Aviator Selection 1919 - 1977. Naval Aerospace Research Laboratory, Naval Air Station, Pensacola, Florida. Special Report 77-2.

Povenmire, H. K. , Alvares, K. M., & Damos, D. L. Observer-observer flight check reliability. Savoy, Ill.: University of Illinois at Urbana-Champaign, Institute of Aviation, Aviation Research Laboratory, Technical Report LF 20-2. 1970.

Biographical Sketch

Gavin Lidderdale is a Principal Psychologist in the Department of the Chief Scientist (RAF), currently working at the Operations Training Division of the USAF Human Resources Laboratory, Williams AFB, Arizona. Whilst in England, he was responsible for the development and validation of the RAF Flying Grading Trial.

Christopher Elshaw is a Senior Psychologist in the Research Branch, Headquarters RAF Support Command. He is an Advisory Member of the RAF Flying Selection Squadron disposal board and is responsible for validation of the selection process.

Lipscomb, M. Suzanne, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas. (Tues. P.M.)

Gender Differences in the Aircraft Maintenance Career Field

Recently, the Air Force has experienced a substantial increase in the number of women working in the non-traditional mechanics career areas. To investigate any differences in the job expectations, experiences, and attitudes of males and females working as Air Force aircraft mechanics, a questionnaire was administered to 800 women and 1600 men in the aircraft maintenance career field. Items were included in the questionnaire to assess why the individual entered the career field, past interest and experience in mechanics, expectations about the job before entering the career field, experiences since entering the career field, and attitudes toward the the job, the Air Force and reenlistment. The responses to the questionnaire were analyzed for overall male/female differences, male/female differences within job types in the career field and the relationships among different experiential factors and attitudes, with specific emphasis on reenlistment intent. Results are discussed in terms of gender differences, job type differences and with respect to retention probabilities.

Gender Differences in the Aircraft Maintenance Career Field

M. Suzanne Lipscomb

Air Force Human Resources Laboratory
Brooks Air Force Base, TX 78235

Background

During the past decade, the number of women serving in the Air Force has increased from approximately 12,000 to more than 60,000. This has taken place while the total active duty force strength has declined by approximately 30%. There are now over 20,000 women working in job specialties which were once considered traditionally male jobs, reflecting a substantial shift in the distribution of sexes across jobs. For example, the Air Force currently has more women working in aircraft maintenance than in personnel specialties.

The substantial increase in numbers of women and their assignment into many traditionally all-male career fields has made it increasingly important for the Air Force to have detailed management information concerning the characteristics of successful and unsuccessful women, the expectations and attitudes of women, and the performance and utilization patterns of women in the operational environment.

While women are assigned to many different technical specialties, there are at present 1,700 women in the Tactical Aircraft Maintenance and Airlift/Bombardment Aircraft Maintenance specialties, comprising approximately 4% of the total population of these two specialties. The large number of women in these traditionally male career fields provides an excellent medium for the comprehensive study of women working in non-traditional areas. Bergmann & Christal (1978) reported a preliminary study investigating the utilization of women in the Aircraft Maintenance career field. Using occupational data collected by the Air Force Occupational Measurement Center (AFOMC), a job-type analysis and an analysis of aptitude distributions were conducted. Even though the data had not been gathered to investigate the utilization patterns of women and, the sample of females was small, the study did produce some interesting findings.

Results from this study indicated that there were significant differences in task assignment as a function of gender. Within the same specialty, a higher percentage of the males were found to be doing actual maintenance tasks while a higher percentage of females were doing support tasks. The data also suggested that during the first enlistment there was a movement of individuals from maintenance to support tasks. However, the extent of this movement appeared to be much larger for females than males. While few differences were found in the tasks performed by males and females working in maintenance functions, differences were found in the tasks performed by men and women working in support functions. Women spent more of their time than did men performing administrative or clerical tasks. These tasks were rated by supervisors as being more difficult than those tasks performed by men in either maintenance or support jobs. Finally, no significant differences were found in job attitudes of males and females in either support or maintenance functions.

The analysis of aptitude distributions conducted in this study suggested that the mechanical aptitude requirements, used for initial classification and assignment and historically predictive of success for males, might not be appropriate for females. Many of the women in the sample did not qualify for entrance into the career field with their scores on the mechanical aptitude test but were admitted into the career field during the period when entry was allowed based on either mechanical or electronic aptitude scores. Nevertheless these individuals had graduated from technical training school and were functioning in the career field and performing maintenance jobs. However, no data were available on their performance levels or on individuals no longer in the specialty.

The findings of this exploratory research indentified the necessity of obtaining data from a larger and more representative sample of males and females in these specialties. These data would include information on the individuals entering the career field, those leaving the career field, and the jobs, tasks, attitudes, and performance levels of those currently in the career field. Information obtained in these areas would provide Air Force management with a comprehensive picture of the utilization of both women and men in the aircraft maintenance career field and allow planning to insure the optimum return on personnel investments. The Manpower and Personnel Division, Air Force Human Resources Laboratory (AFHRL), has initiated a multi-year research program designed to assess: 1) the characteristics of the aircraft maintenance career field input population; 2) the characteristics of those leaving this career field; 3) the on-the-job utilization patterns within the career field; 4) job expectations and attitudes of career field incumbents; and 5) the on-the-job performance of personnel in this career field. This paper will presents the preliminary findings of this research program concerning the job expectations, experiences and attitudes of men and women in the Aircraft Maintenance career field. Investigation in the other areas of interest is continuing and results will be reported at a future time.

Method

Survey Construction

In April 1980, an occupational attitude and experience survey was administered in conjunction with the routine occupational survey of the Aircraft Maintenance career field conducted by the Air Force Occupational Measurement Center. The survey battery contained a background section in which job incumbents provided demographic information about themselves, their job satisfaction, reenlistment intent, TDY experience, shift work and equipment useage. This section was followed by a task list covering 23 duties and 1045 tasks to assess the relative time spent on tasks performed by the respondent. The task list was followed by a final section addressing job expectations, attitudes, and experience. Responses to the task list are undergoing a separate analysis which will be reported at a later date.

The job expectations and experience items were designed to identify why the incumbent entered the career field, their past interest and experience in the mechanical area, their expectations about the job before entering the career field, and their experiences since entering the career field. It is important to know if men and women enter the career field for the same or different reasons, have differing expectations, have different experiences on the job and how these factors effect such things as attitudes and intention to reenlist. The first items in the questionnaire were concerned with the reasons why the respondent entered the Air Force, and the reasons why they entered the aircraft maintenance career field. Respondents were also asked whether or not aircraft maintenance was their first career field choice and what their career plans were when entering the Air Force. These questions were raised by a study conducted in 1974 in which over half of the women in a sample of technical school trainees reported they had chosen aircraft maintenance because it was the only field open when they enlisted (Longridge, 1974). Also relating to attitudes held before entering the career field, respondents were asked about their prior experience and interest in the mechanical area. As an indication of current interest in the area, they were asked if they would try to find a job in a mechanical field if they were to leave the Air Force.

In order to examine the expectations held by men and women before they entered the career field, a series of questions were asked concerning how their work compared to what they had expected. The items covered technical difficulty, physical strength requirements, workload, and environment. These same areas were also covered by questions relating to how the respondents' job had changed since entering the career field. A series of questions concerning job satisfaction were asked followed by questions about the amount of assistance required on their job and specific problems with technical tasks and physical requirements. The final items in the questionnaire concerned supervisor attitudes and a section concerning the male/female composition of the work group and the perceived effectiveness of that work group.

Data Collection

The survey battery was administered in the field to 100 percent of the women in the career field who were available to complete the battery and two men in the career field for every woman surveyed. Women and men were matched on length of military service to eliminate any systematic bias due the fact that the average length of military service of males in the career field, as a whole, is longer than the average length of military service of the women in career field.

Data Analysis

The responses to this questionnaire were analyzed to identify statistically significant male/female differences. Most items had options reflecting an underlying continuum of possible responses and significant gender differences among these items were identified through t ratios computed using the Bonferroni technique for multiple comparisons. In addition to identifying items having significant gender effects, the strength of these effects was also evaluated in terms of the amount of variance items accounted for which discriminated between genders through a stepwise discriminant analysis algorithm. For those few items having categorical response options, a chi-square analysis was performed to identify significant gender differences.

RESULTS

An overall description of the sample obtained for analysis is shown in Table 1. Again, the intent of the sampling strategy was to eliminate any systematic bias resulting from differences in length of military service between males and females. Although there are some minor differences between the female and male sub-samples, overall, the men and women appear well matched. The only difference worth noting is that women report being in their job a shorter length of time than the men. This is not to be confused with time in service or in the career field.

Table 1. Sample Characteristics

Item	Women (n=797)		Men (n=1594)	
	Mean	SD	Mean	SD
1. Age	24.70	3.39	23.76	2.76
2. Years of Education	12.43	.94	12.20	.70
3. Months of Military Service	36.71	20.07	36.71	20.07
4. Military Grade	3.57	.87	3.50	.81
5. Months in Career Field	33.93	20.09	33.77	19.62
6. Months in Job	16.02	12.47	19.88	13.62
7. No. Supervised	.44	1.25	.72	1.93

In Table 2 are shown the results of the chi-square analyses of the questionnaire items having categorical response options. Statistically significant ($p < .001$) gender differences were found on all these items. As shown in Table 2, the reason most given for entering the aircraft maintenance career field by both males and females was that it was the area of strongest interest. However, differences can be noted in all response categories. Similar to the findings of the Longridge study, females were more likely to have been persuaded by a recruiter or to have entered the career field because it was the only one open to them. Differences can also be noted in the reasons given for entering the Air Force. A higher percentage of males than females report entering the AF to gain training for use in civilian life or to serve their country. Females were more likely than males to report entering the Air Force for college or educational benefits or for adventure, excitement and travel. When asked about their career plans upon entering the career field, a larger percentage of females than males reported that aircraft maintenance was not their first career field choice. Similarly, a higher percentage of the females than males planned to crosstrain out of the specialty. A higher percentage of the female sample attended resident technical training and females were more likely to work the day shift as opposed to other shifts. No differences were found as to whether supervisors were male or female, military or civilian.

Table 2. Chi-square Analyses

Item	% Females (n=797)	% Males (n=1594)	Chi-Square
1. What was the main reason you entered the aircraft maintenance (43XXX) career field?			190.69*
1. Only career field open when I enlisted	25.41	13.17	
2. Previous civilian experience in the field	.51	2.49	
3. Training for use in civilian life	3.68	15.66	
4. Area of strongest interest	33.55	47.19	
5. Persuaded by recruiter	28.97	17.20	
6. Other	7.88	4.28	
2. What was the main reason you enlisted in the Air Force?			147.43*
1. Training for use in civilian life	10.76	28.63	
2. College/educational benefits	31.63	21.60	
3. Adventure/excitement/travel	25.33	16.61	
4. Influenced by someone in the military	6.79	6.45	
5. To get away from a personal problem	5.25	4.60	
6. Couldn't find any other job	5.89	6.84	
7. To serve my country	6.40	10.93	
8. Can earn more money in the Air Force than as a civilian	1.92	.32	
9. Other	6.02	4.03	
3. In your present job, which of the following most closely describes the schedule you normally work?			43.40*
1. Day shift (such as 0800-1600)	55.96	41.61	
2. Other	44.04	58.39	
4. What were your career plans when you entered the aircraft maintenance (43XXX) career field?			36.67*
1. I planned to stay in the 43XXX career field for at least my first-term of enlistment	49.43	57.77	
2. I planned to cross train out as soon as possible	20.88	11.74	
3. I had no plans	29.69	30.49	
5. Was aircraft maintenance (43XXX) your first career field choice?			28.61*
1. Yes	60.41	71.46	
2. No	39.59	28.54	

Table 2. Chi-square Analyses (Cont'd)

Item	% Females (n=797)	% Males (n=1594)	Chi-Square
6. How were you assigned to your present career ladder? Blacken only one circle on this line.			25.89*
1. Completed resident technical training	80.05	72.22	
2. Reclassified without completing technical training or OJT	.91	1.54	
3. Directed duty assignment (DDA) from basic training to OJT without bypass test	3.24	6.82	
4. DDA from basic training by bypass test	.26	.96	
5. Converted from another AF specialty without training by classification board action	.52	1.35	
6. Retrained from another specialty	1.04	1.16	
7. Reenlisted after prior service in USAF or from another branch of service	1.17	1.54	
8. Not assigned to my career ladder by any of the above methods	12.82	14.41	
7. Is your immediate supervisor male or female?			.37
1. Male	98.11	98.54	
2. Female	1.89	1.46	
8. Is your immediate supervisor military or civilian?			.60
1. Civilian	3.91	4.70	
2. Military	96.09	95.30	

*Significant at .01 level.

A total of 54 questionnaire and background items with continuous response options were analyzed for male/female differences. T-values were computed to identify significant differences between male and female sample means and a Bonferroni adjustment for multiple comparisons was applied in the test for significance. As shown in Table 3, 25 items were found to be significant at the .01 level.

To determine practical significance, the same 54 items were entered into a stepwise discriminant function analysis with gender as the dependent variable. A total of 13 variables were identified as significantly differentiating between men and women at the .001 level, and accounted for 38% of the variance discriminating between genders. The first six variables entering into the discriminant function accounted for 34% of the total gender variance. However, it was found that variables entering the stepwise algorithm after the sixth item each accounted for less than 1% of the gender variance. Therefore, it was felt that only the first six variables merited discussion.

As shown in Table 4, the item asking "If you left the Air Force, would you try to find a job in the mechanical area?" was the best single discriminator between males and females accounting for approximately 18% of the gender variance. Females were found significantly less likely to attempt to find a mechanical job in the civilian sector after separation than were the males. Other significant discriminators indicated that women reported less prior experience than males in the mechanical area and females found that their jobs required somewhat more physical strength than they expected, whereas males reported their jobs required slightly less physical strength than expected. Women reported working in duty sections with a slightly higher proportion of females than do males and were less inclined to leave the Air Force than were the males. When asked how much confidence their supervisors had in them initially, women reported perceiving less confidence than males. It is interesting to note that when asked how much confidence their supervisors have in them now, there was virtually no difference in the amount of confidence males and females reported perceiving. It is also of interest to note that that on many items in which differences might have been expected, such as the amount of assistance required, and in amount of TDY, meaningful gender differences were not found. As in the Bergmann and Christal study, differences in job attitudes were not found to have practical significance.

CONCLUSIONS

Overall, the analysis of the job expectations, experiences, and attitude data indicate that some significant differences do exist between males and females in their reasons for entering the Air Force and the Aircraft Maintenance career field, in their previous mechanical experience and their plans for civilian work.

Differences were also indicated in the area of expectations about the amount of strength required on the job, initial supervisor confidence and desire to leave the Air Force. However, differences were not found to be gender specific in other areas of expectations, experiences and attitudes. Overall, satisfaction with the Air Force and their jobs, current supervisory confidence, job difficulties and job changes were not found to have discriminative gender significance.

Table 3. Questionnaire Items: Summary Statistics

Item	Females (n=797) Mean	Females (n=797) SD	Males (n=1594) Mean	Males (n=1594) SD	t ratio[a]
1. Present Grade 1. E1-AB 6. E6-TSgt 2. E2-AMN 7. E7-MSgt 3. E3-A1C 8. E8-SMSgt 4. E4-Sgt 9. E9-CMSgt 5. E5-SSgt	3.57	.87	3.50	.81	1.88
2. Total Time in Career Field (months)	33.93	20.09	33.77	19.62	.19
3. Time in Present Job (months)	16.02	12.47	19.88	13.62	-6.79*
4. Circle the highest school grade or college/ university year completed (include equal level, like GED, but not special training, like vocational, outside regular school) 01 07 13 02 08 14 03 09 15 04 10 16 05 11 17 06 12 18	12.43	.94	12.20	.70	6.16*
5. In your present job, which of the following most closely describes the schedule you normally work? 1. Day shift (such as 0800-1600) 2. Other	1.44	.49	1.58	.49	b
6. For how many airmen and civilian are you the immediate supervisor? (Include only those who report directly to you)	.442	1.25	.719	1.93	-3.64
7. Over the last six months, how many days have you been TDY? 1. None 6. 91-120 Days 2. 1-14 Days 7. 121-150 Days 3. 15-30 Days 8. 151-180 Days 4. 31-60 Days 9. 181 Days or more 5. 61-90 Days	1.48	.98	1.79	1.34	-6.36*
8. Over the last year, how many times have you been TDY? 1. None 4. 11-15 Times 7. 26-30 Times 2. 1-5 Times 5. 16-20 Times 8. 31-35 Times 3. 6-10 Times 6. 21-25 Times 9. 36 Times or more	1.50	.89	1.61	.94	-2.95

Table 3. Questionnaire Items: Summary Statistics (Cont'd)

Item	Females (n=797) Mean	SD	Males (n=1594) Mean	SD
9. Was aircraft maintenance (43XXX) your first career field choice? 1. Yes 2. No	1.40	.49	1.28	.45
10. How similar is your assigned career field to your preferred career field? 1. I was assigned to my preferred career field. 2. Very similar 3. Somewhat similar 4. Not very similar 5. Not at all similar	3.07	1.84	2.38	1.63
11. How difficult was aircraft maintenance technical training for you? 1. Very difficult 2. Somewhat difficult 3. Fairly easy 4. Very easy	2.84	.84	3.24	.69
12. Before you entered aircraft maintenance technical training, how much mechanical experience did you have.? 1. Considerable experience 2. Some experience 3. Little experience 4. No experience	3.09	.91	2.13	1.01
13. Before entering aircraft maintenance technical training, how much interest did you have in mechanics? 1. Considerable interest 2. Some interest 3. Little interest 4. No interest	2.32	1.04	1.50	.75
14. If you left the Air Force, would you try to find a job in the mechanical area? 1. Definitely 2. Probably 3. Not sure 4. Probably not 5. Definitely not	3.57	1.29	2.26	1.20
15. How do you find your job? 1. Extremely Dull 2. Very Dull 3. Fairly Dull 4. So-So 5. Fairly Interesting 6. Very Interesting 7. Extremely Interesting	4.64	1.30	4.63	1.35

Table 3. Questionnaire Items: Summary Statistics (Cont'd)

Item	Females (n=797) Mean	SD	Males (n=1594) Mean	SD	t ratio[a]
16. Is your work more or less difficult technically than you expected before entering the 431X1/X2 AFSC? 1. Much more difficult 2. Slightly more difficult 3. It is what I expected 4. Slightly less difficult 5. Much less difficult	3.46	1.14	3.78	1.05	-6.56*
17. Do you do more or less work where you get dirty than you expected before entering the 431X1/X2 AFSC: 1. Much more 2. Slightly more 3. It is what I expected 4. Slightly less 5. Much less	2.93	1.15	2.90	1.08	.73
18. Does your work require more or less physical strength than you expected before entering the 431X1/X2 AFSC? 1. Much more strength 2. Slightly more strength 3. It is what I expected 4. Slightly less strength 5. Much less strength	2.66	1.14	3.22	.94	-11.94*
19. Is your workload heavier or lighter than you expected before entering the 431X1/X2 AFSC? 1. Much heavier 2. Slightly heavier 3. It is what I expected 4. Slightly lighter 5. Much lighter	2.95	1.12	3.04	1.14	-1.86
20. Do you spend more or less time working under unfavorable conditions (heat, cold, bad weather, etc.) than you expected before entering the 431X1/X2 AFSC? 1. Much more 2. Slightly more 3. It is what I expected 4. Slightly less 5. Much less	2.44	1.18	2.34	1.13	1.84

Table 3. Questionnaire Items: Summary Statistics (Cont'd)

Item	Females (n=797)		Males (n=1594)		
	Mean	SD	Mean	SD	t rat[illegible]
21. Do you now receive more or less supervision than when you first began working in the 431X1/X2 AFSC? 1. Much more 2. Slightly more 3. It is what I expected 4. Slightly less 5. Much less	4.32	.96	4.22	1.05	2.[illegible]
22. Do you now have more or less supervisory responsibility than when you first began working in the 431X1/X2 AFSC? 1. I am not a supervisor 2. Much more 3. Slightly more 4. No change 5. Slightly less 6. Much less	3.14	.96	2.93	.87	4.[illegible]
23. Do you do more or less clerical/administrative work than when you first began working in the 431X1/X2 AFSC? 1. Much more 2. Slightly more 3. No change 4. Slightly less 5. Much less	2.51	1.30	2.87	1.12	-6.7[illegible]
24. Do you now do more or less work where you get dirty than when you first began working in the 431X1/X2 AFSC? 1. Much more 2. Slightly more 3. No change 4. Slightly less 5. Much less	3.36	1.39	2.91	1.17	7.7[illegible]
25. Does your work now require more or less physical strength than when you first began working in the 431X1/X2 AFSC? 1. Much more 2. Slightly more 3. No change 4. Slightly less 5. Much less	3.36	1.26	3.10	.96	5.2[illegible]

Table 3. Questionnaire Items: Summary Statistics (Cont'd)

Item	Females (n=797) Mean	Females (n=797) SD	Males (n=1594) Mean	Males (n=1594) SD	t ratio[a]
26. Is your workload now heavier or lighter than when you first began working in the 431X1/X2 AFSC? 1. Much heavier 2. Slightly heavier 3. No change 4. Slightly lighter 5. Much lighter	2.77	1.28	2.60	1.17	3.28
27. Do you now spend more or less time working under unfavorable conditions (heat, cold, bad weather, etc.) than when you first began working in the 431X1/X2 AFSC? 1. Much more 2. Slightly more 3. No change 4. Slightly less 5. Much less	3.24	1.31	2.82	1.11	7.70*
28. Are you satisfied with the Air Force in general? 1. Very satisfied 2. Fairly satisfied 3. Neither satisfied or dissatisfied 4. Somewhat dissatisfied 5. Very dissatisfied	2.86	1.25	3.15	1.26	-5.36*
29. How satisfied are you with your present Air Force job? 1. Very satisfied 2. Fairly satisfied 3. Neither satisfied or dissatisfied 4. Somewhat dissatisfied 5. Very dissatisfied	3.06	1.37	2.90	1.33	2.61
30. How much useful skill and experience are you gaining from your job? 1. A very large amount 2. A large amount 3. Some 4. Very little 5. None	3.11	1.14	2.78	1.14	6.66*
31. How does your job utilize your talents? 1. Not at all 5. Very well 2. Very little 6. Excellently 3. Fairly well 7. Perfectly 4. Quite well	3.15	1.34	3.37	1.43	-3.68

Table 3. Questionnaire Items: Summary Statistics (Cont'd)

Item	Females (n=797) Mean	Females (n=797) SD	Males (n=1594) Mean	Males (n=1594) SD	t
32. How does your job utilize your training? 1. Not at all 5. Very well 2. Very little 6. Excellently 3. Fairly well 7. Perfectly 4. Quite well	3.33	1.44	3.66	1.45	
33. How satisfied are you with the sense of accomplishment you gain from your work? 1. Extremely Dissatisfied 5. Slightly Satisfied 2. Very Dissatisfied 6. Very Satisfied 3. Slightly Dissatisfied 7. Extremely Satisfied 4. Neither Satisfied Nor Dissatisfied	4.43	1.72	4.54	1.64	
34. If you had the chance, would you change to another career field? 1. Definitely would 2. Probably would 3. Not sure 4. Probably would not 5. Definitely would not	2.18	1.24	2.39	1.27	
35. If you had the chance, would you leave the Air Force? 1. Definitely would 2. Probably would 3. Not sure 4. Probably would not 5. Definitely would not	2.97	1.27	2.65	1.26	
36. Do you plan to reenlist? 1. No 2. Uncertain, Probably No 3. Uncertain, Probably Yes 4. Yes	2.32	1.10	2.25	1.06	
37. How many of the tasks in your job can you perform without technical assistance from another person? 1. A very large number 2. A large number 3. Some 4. Very few 5. None	1.88	.819	1.78	.78	
38. How often do coworkers volunteer to give you technical assistance? 1. Very often 2. Often 3. Sometimes 4. Rarely 5. Never	2.76	1.07	2.78	1.02	

Table 3. Questionnaire Items: Summary Statistics (Cont'd)

Item	Females (n=797) Mean	SD	Males (n=1594) Mean	SD	t ratio[a]
39. How often do you have to ask for technical assistance? 1. Very often 2. Often 3. Sometimes 4. Rarely 5. Never	3.25	.75	3.26	.78	-.29
40. How many of the physical tasks (lifting, moving equipment, reaching, etc.) in your specialty can you perform without assistance from another person? 1. All 2. Most 3. Some 4. Few 5. None	2.40	.87	2.26	.74	3.98*
41. How often do coworkers volunteer to help you with physical tasks (lifting, moving equipment, reaching, etc.)? 1. Very often 2. Often 3. Sometimes 4. Rarely 5. Never	2.52	1.05	2.51	.99	.27
42. How often do you have to ask for help doing physical tasks? (Lifting, moving equipment, reaching, etc.) 1. Very often 2. Often 3. Sometimes 4. Rarely 5. Never	3.27	.83	3.22	.84	1.34
43. How many of the tools required in your job are difficult to use because of their size or bulkiness? 1. A very large number 2. A large number 3. Some 4. Very few 5. None 6. Not applicable	3.90	.91	3.76	.97	3.12

Table 3. Questionnaire Items: Summary Statistics (Cont'd)

Item	Females (n=797)		Males (n=1594)		
	Mean	SD	Mean	SD	t r
44. How many of the tools required in your job are difficult to use because they require too much strength? 1. A very large number 2. A large number 3. Some 4. Very few 5. None 6. Not applicable	4.04	.85	4.35	.76	-[illegible]
45. How many of the tools required in your job are difficult to use because it is hard to understand how to operate them? 1. A very large number 2. A large number 3. Some 4. Very few 5. None 6. Not applicable	4.50	.70	4.53	.69	-[illegible]
46. Is your immediate supervisor male or female? 1. Male 2. Female	1.02	.14	1.01	.12	
47. Is your immediate supervisor military or civilian? 1. Civilian 2. Military	1.96	.19	1.95	.21	
48. How does your supervisor judge your work? 1. Very leniently 2. Somewhat leniently 3. In a fair way 4. Somewhat harshly 5. Very harshly	2.97	.62	2.92	.68	[illegible]
49. How often does your immediate supervisor give you recognition for a job well done? 1. Very often 2. Often 3. Sometimes 4. Rarely 5. Never	2.96	1.12	3.08	1.11	-[illegible]

Table 3. Questionnaire Items: Summary Statistics (Cont'd)

Item	Females (n=797) Mean	SD	Males (n=1594) Mean	SD	t ratio[a]
50. When you first began this assignment, how much confidence did your supervisor have that you could do your job well, compared to your fellow workers with the same amount of experience? 1. A very large amount 2. A large amount 3. Some 4. Very little 5. None 6. Don't know	2.86	1.09	2.45	.92	8.30*
51. How much confidence does your supervisor have that you can do your job well now, compared to your fellow workers with the same amount of experience? 1. A very large amount 2. A large amount 3. Some 4. Very little 5. None 6. Don't know	1.96	.83	1.96	.81	.03
52. How would you rate the effectiveness of your duty section compared to other 43XXX duty sections? 1. Much more effective 2. More effective 3. About the same 4. Slightly less effective 5. Much less effective	2.68	1.02	2.56	1.11	2.53
53. Approximately how many 431X1/X2 personnel are in your duty section? 1. 0-5 6. 26-30 2. 6-10 7. 31-35 3. 11-15 8. 36-40 4. 16-20 9. 41 or more 5. 21-25	5.68	3.08	6.17	2.75	-3.75*
54. Approximately what is the proportion of female and male 431X1/X2 personnel in your duty section? 1. 0% female/100% male 2. 5% female/95% malle 3. 10% female/90% male 4. 15% female/85% male 5. 20% female/80% male 6. 30% female/70% male 7. 40% female/60% male 8. 50% or more female/50% or less male	2.39	1.14	1.94	.86	9.92*

[a]Bonferroni $T_{crit} = 3.75$, $p = .01$

[b]These items are not amenable to T-test comparisons. See Table 2.
$*p < .01$

Table 4. Summary of Significant Gender Discriminators

Item	Females (n=521) Mean	Females (n=521) SD	Males (n=1137) Mean	Males (n=1137) SD	Unique Contribution
1. If you left the Air Force, would you try to find a job in the mechanical area? 1. Definitely 2. Probably 3. Not sure 4. Probably not 5. Definitely not	3.57	1.29	2.26	1.20	18%
2. Before you entered aircraft maintenance technical training, how much mechanical experience did you have? 1. Considerable experience 2. Some experience 3. Little experience 4. No experience	3.09	.91	2.13	1.01	7%
3. Does your work require more or less physical strength than you expected before entering the 431X1/X2 AFSC? 1. Much more strength 2. Slightly more strength 3. It is what I expected 4. Slightly less strength 5. Much less strength	2.66	1.14	3.22	.94	4%
4. If you had the chance, would you leave the Air Force? 1. Definitely would 2. Probably would 3. Not sure 4. Probably would not 5. Definitely would not	2.39	1.14	1.94	.86	2%
5. Approximately what is the proportion of female and male 431X1/X2 personnel in your duty section? 1. 0% female/100% male 2. 5% female/95% male 3. 10% female/90% male 4. 15% female/85% male 5. 20% female/80% male 6. 30% female/70% male 7. 40% female/60% male 8. 50% or more female/50% or less male	2.97	1.27	2.65	1.26	2%
6. When you first began this assignment, how much confidence did your supervisor have that you could do your job well, compared to your fellow workers with the same amount of experience? 1. A very large amount 2. A large amount 3. Some 4. Very little 5. None 6. Don't know	2.86	1.09	2.45	.92	1%

Perhaps the information to come from these data which should be of most interest to Air Force managers is that a greater percentage (54.38%) of the women than men (30.37%) in this study entered the aircraft maintenance career field because it was the only area open or they were persuaded by the recruiter. Similarly, approximately 21% of the women compared to 12% of the men planned to crosstrain out at the time of entry. While these differences in reason for entering the career field and career plans have no corresponding difference in job attitudes and reenlistment intent, it is not yet known the impact of these differences in other areas such as crosstraining rates and performance levels.

It is also of interest that the women in the sample reported perceiving less confidence initially from their supervisors than the men. This, as well as the women expecting the job to require less strength than it did, could result in a longer time being required for the women to adjust to the job initially. These misconceptions on the part of the supervisors as well as those entering the career field could be overcome through the dissemenation of more accurate information and efforts to dispel predjudice. As women report less interest in leaving the Air Force than do men, efforts to improve their utilization and adjustment into the career field should be worthwhile.

Further analysis of these data will be conducted using the results of the currently ongoing job analysis. Males and females can then be grouped by job type and differences between males and females in the same job type and across job types can be analyzed to investigate the effect of the type of work performed upon these areas of interest. This, along with the investigation of the input and exit populations, utilization patterns, and on-the-job performance levels, will give an in-depth view of the career field and the women and men working in it.

References

Bergmann, J.A., & Christal, R.E. Female utilization in non-traditional areas. Proceedings, 20th Annual Conference of the Military Testing Association, Oklahoma City, October - November 1978, 444-463.

Longridge, T.M., Jr. Training of Enlisted Women as Jet Aircraft Maintenance Specialists. (Sheppard PR 74-4). Sheppard Air Force Base, Texas: Training Research Applications Branch, USAF School of Applied Aerospace Sciences, August 1974. (AD/A - 006434).

AD P001346

Loo, Robert, Canadian Forces Personnel Applied Research Unit, Willowdale, Ontario, Canada. (Tues. P.M.)

Utility of the Field Dependence Dimension in Military Settings

The field dependence - field independence dimension was defined and its relationships with various personality dimensions, social behaviors, and problem-solving styles were identified. In the military context, the utility of the field dependence construct and of various measures of field dependence were critically reviewed. Empirical studies which used military samples or which examined areas directly related to the military, for example, map reading skills, target detection, and computer skills were examined. Then conceptual and methodological issues surrounding the field dependence construct were discussed. These issues focused on the relationships between field dependence and intelligence, sex differences in field dependence, relationships among various measures of field dependence, components underlying performance on field dependence tests, and design problems with much of the research in this area. Finally, a set of guidelines was presented for the future use of the construct and its measures in military settings. These guidelines focused on, among other points, the importance of selecting appropriate measures of the construct, the use of multivariate as well as univariate designs, and controls for various potential sources of confounds.

UTILITY OF THE FIELD DEPENDENCE DIMENSION IN MILITARY SETTINGS

Robert Loo Ph.D.
Canadian Forces Personnel Applied Research Unit
4900 Yonge Street
Willowdale, Ontario

INTRODUCTION

Beginning with visual perception studies using the Rod-and-frame Test, Witkin and his associates (e.g., Witkin, Lewis, Hertzman, Machover, Meissner & Wapner, 1954) found that subjects could be differentiated along a continuum based on their ability to separate a given "figure" from "ground" in the visual field. They labelled this continuum perceptual field-dependence-independence. Field dependence (FD) can be described as a "fused" or "global" way of perceiving; perception is dominated by the overall organization of the (visual) field. Field independence (FI), on the other hand, can be described as an articulated way of perceiving; parts of the field are perceived as discrete from the organized background.

In the 30-odd years since its introduction into the scientific literature, the field dependence dimension has been extensively examined. Before his unfortunate death in 1979, Witkin (Witkin & Goodenough, 1977; Witkin, Goodenough & Oltman, 1979) published two papers which are excellent overviews on the status of the field dependence construct. In one paper, Witkin & Goodenough (1977) presented a comprehensive examination of the relationships between field dependence and interpersonal behavior. Essentially, they reported that FD persons have a strong interpersonal orientation in terms of being emotionally open, gravitating toward social situations, and prefering to be physically close to others; whereas, FI persons have an impersonal orientation in terms of not being very interested in others, showing both physical and psychological distancing from people, and showing a preference for nonsocial situations. In the other paper, Witkin (Witkin et al., 1979) updated the status of his psychological differentiation construct, a high-order construct of which field dependence is the most studied lower-order construct. An important focus of this paper, as compared with Witkin's previous overviews, was the attention given to neurophysiological research published during the 1970's on the relationships between field dependence and, say, lateral specialization of the hemispheres or patterns of electroencephalogram (EEG) recordings.

MEASURES OF FD

The pioneering research used the Rod-and-Frame Test (RFT), a test in which the subject, while seated in a darkened room, must adjust to the upright a tilted luminous rod centered within a tilted luminous frame. Since those pioneering days, Witkin and his associates have developed several other tests of field dependence. Several tests are variations of the RFT; for example, the portable RFT, the Tilting-Room-Tilting-Chair Test, the Room-Adjustment Test, and the Body-Adjustment Test. However, another line of test development produced the Embedded Figures Test (EFT) and a group-administered version of the EFT. These embedded-figures tests are more practical in the sense that they do not require cumbersome equipment, special rooms or periods of dark adaptation. In addition, the Group Embedded Figures Test (GEFT) permits group testing whereas all the other tests mentioned require individual testing, a time consuming and costly way of gathering data.

Numerous studies have examined the interrelationships among the various measures of field dependence. Arbuthnot (1972) reviewed 40 such studies which reported a total of 122 correlations between field dependence measures. He noted, for example, that the mean correlation was 0.54 for the 21 correlations reported between scores on the RFT and Witkin's original 24-item EFT.

FD AND OTHER VARIABLES

There has been much interest in relating field dependence to variables in the personality and attitudes domains among others. Table 1 identifies some findings reported in the literature over the past 25 years. One caveat in examining the table is that for each area there are studies which have failed to obtain the described relationships and even studies which have obtained contradictory results. However, there appears to be sufficient evidence to support the descriptions in Table 1. The two papers by Witkin (Witkin & Goodenough, 1977; Witkin et al., 1979) provide references for these various relationships.

Table 2 presents some applications of field dependence which are military based or directly applicable to the military. Fine (e.g., 1980), at the U.S. Army Research Institute of Environmental Medicine, has demonstrated the utility of the field dependence construct in studying (visual) sensation and perception in a military context where U.S. soldiers served as participants. Field dependence has also been used as one variable in describing human operator performance in man-machine interfaces; for example, Howard (1979) has shown the usefulness of field dependence in predicting console operator performance on the Patriot weapon system. Moving from performance in man-machine systems to performance "in the field", Belletti and Anthony (1979) have shown that field independent persons on an artillery forward observers course performed better than field dependent persons. Studies have also shown the utility of field dependence in examining pilot performance (e.g., Cullen, 1969; Kennedy, 1972), vehicle driver performance (e.g., Loo, 1978a), and performance in vehicle simulators (e.g., Barrett et al., 1969).

Stasz and Thorndyke (1980) have shown that field dependence is related to map reading skills where such skills are relevant to many areas in the military. Loo (1981) has suggested that field dependence may be a useful construct in selecting and in understanding performance in tactical command, control, and communication (C^3) systems. Finally, work by Fine (1972) with military samples and by Loo (1978b) with civilian samples has shown relationships between field dependence and symptoms of psychopathology, namely, neuroticism.

CRITIQUE OF FIELD DEPENDENCE

Many have taken up the field dependence construct in their research since Witkin's early work; however, several researchers (e.g., Arbuthnot, 1972; Vernon, 1972; Wachtel, 1972) have expressed concern over the validity of some claims made by Witkin and over the methodology used in many studies in this area.

Performance Components of FD Measures

It is suggested that, on the basis of determinants of performance, typical measures of field dependence can be clustered into two groups, embedded-figures tests and adjustment test such as the Rod-and-Frame Test. This position differs from Witkin who would claim that these tests have high convergent validity in the measurement of field dependence.

Perhaps the earliest evidence for the two-clusters position espoused in the present paper can be found in Goodenough and Karp's (1961) factor-analytic study. Although they were concerned with the relationship between field dependence and intellectual functioning, their results can be interpreted to suggest that the various field-dependence variables they used do not clearly identify a homogeneous clustering of field-dependence measures. They performed two oblique factor analyses, one on data from a group of 25 boys and 25 girls (Group A) and the other on data from a group of 30 boys (Group B). Factor loadings in the solution for Group A showed that no one factor identified all eight field-dependence variables. However, loadings for the even smaller Group B showed that Factor III identified all five field-dependence variables used with that group. In both oblique solutions, each of the various field-dependence variables typically loaded on more than one factor. Unfortunately, the authors did not report the correlations among the primary factors or second-order analyses. Obviously, oblique primaries indicated that higher-order analyses should have been performed, and if they had been, then the results might have been more meaningful both for their purposes and for those of the present paper.

A second more comprehensive and direct source of evidence for the position is found in Vernon's (1972) large-scale factor-analytic study with adolescents. First, his two Rod-and-frame Test scores intercorrelated higher (0.63) than the two scores did with Embedded-figures Test scores (0.32, 0.40); and secondly, Vernon found that when general intelligence was partialed out of all correlations, only the intercorrelation between the two Rod-and-frame Test scores remained significant (0.53), while the remainder of the correlations, including those between the two Rod-and-frame Test scores and the Embedded-figures Test scores (0.15, 0.18), became nonsignificant. These points suggest that performance on the Rod-and-frame Test is essentially independent of intelligence, whereas performance on the Embedded-figures Test is highly related to intelligence level.

Finally, Arbuthnot (1972) reviewed 40 studies which reported correlations among various field-dependence measures. He noted that the mean correlation between performance on the Rod-and-frame Test and the 24-item Embedded-figures Test from 21 correlations was 0.54, and between the Rod-and-frame Test and the 12-item Embedded-figures Test from nine correlations was 0.37. The magnitude of the correlations indicated the low common variance shared by these two clusters of measures and thus the lack of convergent validity.

COMPONENTS UNDERLYING PERFORMANCE IN EMBEDDED-FIGURES AND ADJUSTMENT TESTS

Several possible components or factors affecting performance on field-dependence tests may be extracted to explain the low covariation between tests in the two clusters, embedded-figures and adjustment tests, and the high covariation within clusters. There seem to be components required for the solving of embedded-figures tests but not for adjustment tests and vice versa. One distinction between the various embedded-figures and adjustment tests such as the Rod-and-frame Test, the portable Rod-and-frame Test, the Tilting-room-tilting-chair Test, Room-adjustment test, and Body-adjustment Test is the requirement in the embedded-figures tests to remember for at least a few seconds the given simple geometric form in order to find it when given the complex form. Secondly, a perceptual-motor component also underlies performance on embedded-figures tests in that the subject is required not only to remember a simple form but also to outline it within the complex form.

A third factor found only in embedded-figures tests is reversible-perspective items. Vojtisek and Magaro (1974) reported that psychiatric patients had greater difficulty solving items with reversible perspective than with other items on embedded-figures tests. On the other hand, Loo (1978c) reported that normal females were at least as successful in solving reversible-perspective items as in solving the remaining items in both individual and group forms of the Embedded-figures Test. However, he also found that when extreme groups were formed based on ease and difficulty of solving reversible-perspective items, greater difficulty was associated with greater sociability and more minor "psychiatric" complaints than was greater ease of solving such items. Finally, the time restraints placed on solution time to items in embedded-figures tests also distinguish these tests from adjustment tests.

At lease one factor or component is required in the adjustment tests but not in the embedded-figures tests. The adjustment tests require that a kinesthetic component be linked with the visuospatial component to achieve high accuracy in these test situations.

Having stressed the components required in one cluster of tests versus the other, the components required in the solution of both clusters of tests are itemized. As in all tests, an optimal level of motivation and arousal (e.g., Oltman, 1964) on the subjects' part is necessary to achieve "best" performance. In addition, the pattern of eye movements is related to performance in both clusters of tests if visual stimuli are used. For example, Blowers and O'Connor (1978) found that with the Rod-and-frame Test, field-independent subjects, unlike field-dependent subjects, showed large-magnitude eye movements and high rates of eye movements.

Additional, although indirect, support for the eye-movements component is found in Baron's (1978) study which investigated the eye movements of 85 children during their television watching. She found that field-independent subjects oriented to the target words faster, had more fixations on target, and had longer fixation durations than did field-dependent subjects.

An extension of the research relating eye movements and field dependence to include eye tracking may prove fruitful. Recently, several groups of researchers (e.g., Holzman et al., 1976; Kuechenmeister et al., 1977) reported impaired eye tracking in various groups of psychotic patients. Essentially, such studies found that in a simple test of smooth-pursuit eye movements, a high proportion of psychotic patients showed imparied performance which was due to velocity arrests. Further investigation (e.g., Holzman et al., 1976) suggested that velocity arrests were due not to voluntary processes but to neurophysiological dysfunctions probably located in the brain stem. However, recent work by Acker and Toone (1978) demonstrated that impaired eye tracking could be induced in normals by the addition of a distracting task. They concluded that contrary to previous research which stressed a neurophysiological deficit, superficial inattention or deficits in selective attention might account for the schizophrenics' poor performances. In any event, research in this area is very active and may prove very fruitful for many areas of psychology.

In addition to eye movements and eye tracking, lateral eye movements as studied in relation to information processing and hemispheric dominance (e.g., Huang and Byrne, 1978) may be of importance in the study of field dependence. Along the same lines, handedness and other laterality indicators are of significant interest in relating cortical organization to the field-dependence dimension (e.g., O'Connor and Shaw, 1978).

The comprehensive examination through multivariate and univariate techniques of the interrelationships involving eye movements, eye tracking, lateral eye movements, laterality, and field-dependence measures in various populations might yield important information on the role of cortical and subcortical structures and processes in the field-dependence dimension.

FIELD-DEPENDENCE MEASURES AND INTELLIGENCE

The controversy over the relationship between field dependence and intelligence is long-standing and unresolved. Findings from studies which examined the relationship tend to indicate that the relationship between scores on performance subtests (Block Design, Object Assembly, Picture Completion) from both the Wechsler Intelligence Scale for Children and Wechsler Adult Intelligence Scale and field-dependence measures is carried by the embedded-figures tests and not the adjustment tests. The rotated factor matrix for Group A reported by Goodenough and Karp (1961) showed that the Room-adjustment Test and Body-adjustment Test loaded highly on factors separate from subtests on the Wechsler Intelligence Scale for Children. In contrast, the rotated factor matrix for Group B showed that all the field-dependence tests and the three subtests, Block Design, Object Assembly, and Picture Completion, loaded on one factor.

Although the focus has been on relating field-dependence measures to performance subtests from intelligence scales, some researchers identified relationships between embedded-figures but not adjustment tests and verbal subtests. Two groups of researchers (Karp and Silberman, 1966; Riley and Denmark, 1974) found that with samples of black subjects, performance on the children and adult forms of the Embedded-figures Test was related to performance on verbal subtests from Wechsler's intelligence scales. More recently, O'Leary et al. (1977) found that field dependence as measured by the group Embedded-figures Test was related to several verbal and performance subtests on the Wechsler-Bellevue Intelligence Scale for both alcoholic and nonalcoholic groups of males. In contrast to these findings, Vernon (1972) noted that when intelligence was held constant the correlation between the two Rod-and-frame Test scores was only slightly attenuated from 0.63 to 0.53, while the correlations between these two scores and the Embedded-figures Test scores were reduced in magnitude by half to nonsignificant levels (0.15, 0.18).

Findings such as these suggest that performance on embedded-figures tests is highly related to performance on both performance and verbal intelligence tests. On the other hand, performance on adjustment tests such as the Rod-and-frame Test is apparently only slightly related to performance on intelligence tests. The differing relationships between embedded-figures and adjustment tests indicate that critics who state that field dependence is highly related to or even the same as intelligence must qualify their statement as applying to embedded-figures tests only (see Fine, 1973, Note 5).

The finding that performance on verbal subtests is related to performance on embedded-figures tests (O'Leary et al., 1977; Riley and Denmark, 1974), combined with the finding that sex differences exist in reported strategies for solving visuospatial tests, suggests a further component, verbal mediation, underlying performance on embedded-figures and possibly, too, adjustment tests. It is suggested that greater field independence as measured by embedded-figures tests may be due, in part, to the effective use of verbal mediation by subjects in their problem-solving approach to test items. Loo and Townsend (1977) found that greater field independence, as measured by the group Embedded-figures Test, was associated with lower impulsivity, specifically slower decision time. This finding provided indirect support for a possible verbal mediation component. Verbal mediation, a time-dependent and reflective behavior, would necessitate low impulsivity and slow decision times.

FUTURE DIRECTIONS FOR APPLICATIONS OF FD

Methodology

Given the number and variety of components underlying performance in the two clusters of tests, embedded-figures and adjustment tests, it is obvious that the understanding of the high-order construct field dependence and of its relationship to other cognitive, personality, and performance dimensions requires the execution of coprehensive rather than limited studies which simply conduct one-way analyses of variance or Pearson correlations. It is recommended that studies employ samples which adequately cover the range of possible scores on field-dependence measures, that multiple measures and scores of field dependence be used as within-subject variables, and that multivariate and multiple regression techniques be considered.

Military Applications

Military applications of the field depencence construct in selection and training may show their worth in the following major areas.

<u>High technology environments</u>. The revolutionary advances in computer hardware and software are penetrating all areas of the military and this trend will continue over the remainder of the century. The potential uses of field dependence in areas such as computer-based management information systems

(e.g., Zmud, 1980) and command, control and communication (C^3) systems (e.g., Loo, 1981) have been identified.

Vehicle operation. The visual-spatial and kinesthetic components associated with field dependence tests and the findings of past research with vehicle driving (e.g., Loo, 1978a) and pilots (e.g., Cullen et al., 1969) certainly supports the continued use of field dependence in this area.

Visual-spatial tasks. Aside from the visual-spatial requirements for vehicle operation as noted above, many military occupations require performance in complex visual-spatial work environments. For example, field dependence may be useful where visual (CRT) displays are used extensively (e.g., radar operators, air traffic controllers, air weapons controllers), where map reading and/or photo interpretation are essential activities, and where visual disembedding is critical such as for forward observers in the artillery and air observers in search and rescue.

Decision making processes. Field dependence, as well as other cognitive styles, may be useful in the area of decision making and decision analysis. Commanders at all levels are decision makers who must deal with information/communications activities under varying conditions of ambiguity, uncertainty, conflict, and ignorance; thus field dependence may be a useful construct in examining decision making (see Loo, 1981; Witkin & Goodenough, 1977).

Drug abuse. Drug abuse in terms of alcohol, cannibis, and other drugs is a concern in the military. The field dependence construct has already shown utility in the study of addictions (e.g., O'Leary et al., 1977) and future research should add further to the literature.

REFERENCES

Acker, W., & Toone, B. Attention, eye tracking and schizophrenia. British Journal of Social and Clinical Psychology, 1978, 17, 173-181.

Arbuthnot, J. Cautionary note on measurement of field independence. Perceptual and Motor Skills, 1972, 35, 479-488.

Baron, L.J. Relating eye movements to reading proficiency, field articulation. and stimulus mode. Paper presented at the meeting of the Canadian Psychological Association, Ottawa, June 1978.

Barrett, G.V., Thornton, C.L. & Cabe, P.A. Relation between embedded figures test performance and simulator behavior. Journal of Applied Psychology, 1969, 53, 253-254.

Belletti, H.E. & Anthony, J.G. The embedded figures test: its relationship to forward observer performance. Paper presented at the meeting of the Military Testing Association, San Antonio, November 1979.

Blowers, G.H. & O'Connor, K.P. Relations of eye movements to errors on the rod-and-frame test. Perceptual and Motor Skills, 1978, 46, 719-725.

Cullen, J.F., Harper, C.R. & Kidera, G.J. Perceptual style differences between airline pilots and engineers. Aerospace Medicine, 1969, 4, 407-408.

Fine, B.J. Field-dependent introvert and neuroticism: Eysenck and Witkin united. Psychological Reports, 1972, 31, 939-956.

Fine, B.J. Field-dependence-independence as "sensitivity" of the nervous system: supportive evidence with color and weight discrimination. Perceptual and Motor Skills, 1973, 37, 287-295.

Fine, B.J. & Kobrick, J.L. Field dependence, practice, and low illumination as related to the Farnsworth-Munsell 100-hue test. Perceptual and Motor Skills, 1980, 51, 1167-1177.

Goodenough, D.R. & Karp, S.A. Field dependence and intellectual functioning. Journal of Abnormal and Social Psychology, 1961, 63, 241-246.

Holzman, P.S., Levy, D.L. & Proctor, L.R. Smooth pursuit eye movements, attention, and schizophrenia. Archives of General Psychiatry, 1976, 33, 1415-1420.

Howard, C.W. Psychological profiles and performance characteristics of tactical console operators. Paper presented at the meeting of the Military Testing Association, San Antonio, November 1979.

Huang, M & Byrne, B. Cognitive style and lateral eye movements. British Journal of Psychology, 1978, 69, 85-90.

Karp, S.A. & Silberman, L. Field dependence, body sophistication, and socioeconomic status. Research Reports, 1966, 1, 1-9, Sinai Hospital, Baltimore.

Kennedy, R.S. The relationship between habituation to vestibular stimulation and vigilance: individual differences and subsidiary problems (Doctoral dissertation, University of Rochester, 1972). Dissertation Abstracts International, 1972, 33, 2374B-2375B.

Kuechenmeister, C.A., Linton, P.H., Mueler, T.V. & White, H.B. Eye tracking in relation to age, sex, and illness. Archives of General Psychiatry, 1977, 34, 578-579.

Loo, R. Individual differences and the perception of traffic signs. Human Factors, 1978a, 20, 65-74.

Loo, R. Relationship of extraversion and field dependence to neuroticism. Psychology, 1978b, 15, 56-66.

Loo, R. Personality dimensions and reversible perspective in embedded figures test. Perceptual and Motor Skills, 1978c, 46, 1016-1018.

Loo, R. Individual differences as human factors in tactical communications systems. Paper presented at meeting of the Human Factors Association of Canada, Toronto, October 1981.

Loo, R. & Townsend, P.J. Components underlying the relation between field dependence and extraversion. Perceptual and Motor Skills, 1977, 45, 528-530.

O'Connor, K.P. & Shaw, J.C. Field dependence, laterality and the EEG. Biological Psychology, 1978, 6, 93-109.

O'Leary, M.R., Donovan, D.M. & Chaney, E.F. The relationship of perceptual field orientation to measures of cognitive functioning and current adaptive abilities in alcoholics and nonalcoholics. Journal of Nervous and Mental Disease, 1977, 165, 275-282.

Oltman, P.K. Field dependence and arousal. Perceptual and Motor Skills, 1964, 19, 441.

Riley, R.T. & Denmark, F.L. Field dependence and measures of intelligence: some reconsiderations. Social Behavior and Personality, 1974, 2, 25-29.

Stasz, C. & Thorndyke, P.W. The influence of visual-spatial ability and study procedures on map learning skill. Rand Corp / N-1501-ONR, 1980.

Vernon, P.E. The distinctiveness of field independence. Journal of Personality, 1972, 40, 366-391.

Vojtisek, J.E. & Magaro, P.A. The two factors present in the embedded figures test and a suggested short form for hospitalized psychiatric patients. Journal of Consulting and Clinical Psychology, 1974, 42, 554-558.

Wachtel, P.L. Field dependence and psychological differentiation: re-examinantion. Perceptual and Motor Skills, 1972, 35, 179-189.

Witkin, H.A. & Goodenough, D.R. Field dependence and interpersonal behavior. Psychological Bulletin, 1977, 84, 661-689.

Witkin, H.A., Goodenough, D.R. & Oltman, P.K. Psychological differentiation: current status. Journal of Personality and Social Psychology, 1979, 37, 1127-1145.

Witkin, H.A., Lewis, H.B., Hetzman, M., Machover, K., Meissner, P.B. & Wagner, S. Personality through perception. Harper: New York, 1954.

Zmud, R.W. An information processing conceptualization of the systematic-heuristic cognitive style. Manuscript from author at Georgia State University, 1980.

TABLE 1

CORRELATES OF FIELD DEPENDENCE-INDEPENDENCE

AREA	GREATER FD	GREATER FI
. Autonomy in social situations	Low	High
. Motivation orientation	Extrinsic	Intrinsic
. Sociability	High	Low
. Impulsivity	High	Low
. Self-reliance	Low	High
. Vocational interests & Career choice	Humanitarian-Helping domains	Theoretic (analytic) - Artistic domains
. Sex differences	Females tend to be more FD	Males tend to be more FI
. Anxiety	High	Low
. Self-reported personal attributes	In general, FI people tend to describe themselves and to be described by others in such terms as friendly, considerate, warm, affectionate, polite, tactful, accommodating, nonevaluative and accepting of others, like people and are liked by others, and make others feel comfortable with them	In contrast, the descriptions of FI people include, among other characteristics, inconsiderate, rude, demanding, ambitious, interested in power, opportunistic, and manipulate people as a means of achieving personal ends
. Neurophysiological Differentiation	Low	High
. Psychopathology	Identity problems, over dependence, inadequate emotional controls	Delusions, outward aggressiveness, overideation...

TABLE 2

APPLICATIONS OF FD

- Sensation & Perception
- Human Operator Performance
  - Tactical console operators
- Forward Observation Performance
- Pilot Performance
- Vehicle Driver Performance
- Performance in Simulators
- Map Reading
- Commanders in C^3 Systems
- Psychopathology

AD P002 347

T & E Rating With Objectively Scored Applicant Appraisal Questionnaires

Thomas J. Lyons Ph.D
Personnel Psychologist

A procedure based on a content validity strategy has been developed for constructing Objectively Scored Applicant Appraisal Questionnaires (OSAAQ) to replace traditional ratings of training and experience (T&E Ratings). There are both technical and administrative problems with traditional T&E Ratings that depend on examiner judgement to evaluate narrative responses to open ended questions against general rating guides or benchmarks. An OSAAQ exam consists of constructed response questions (e.g., multiple choices, check list) with an objective scoring key for evaluating applicant responses. This makes OSAAQ items as appropriate for automated processing and scoring as objective written test items and minimizes administrative costs with large numbers of applicants. Objective item formats and scoring procedure of OSAAQ exams also ensure consistent evaluation of applicant qualifications. General procedures for constructing OSAAQ exams are described with a discussion of technical and administrative issues related to OSAAQ exam development projects for several occupations.

The model described in this paper is designed to improve the efficiency, job relatedness and reliability of unassembled exams used to evaluate applicants for high volume Federal occupations. Unassembled exams refer to a variety of procedures for rating training and experience (T & E ratings) which require raters (usually personnel staffing or occupational specialists) to evaluate narrative information summarized on application and/or supplemental forms. These procedures are typically used to examine for full performance and level jobs and entry level professional and technical jobs requiring specialized training or experience. Although T & E ratings are usually practical to apply with small numbers of applicants, they become costly to administer when the volume of applications is high. It is also difficult to ensure adequate variability and consistency in applicant ratings for traditional T & E ratings that are applied to high volume occupations. These concerns have led to the development of a new unassembled examining methodology for constructing Objectively Scored Applicant Appraisal Questionnaires (OSAAQ) that is practical to use and capable of objectively differentiating among large numbers of applicants on job related factors.

Background

The OSAAQ exam development methodology is used to develop job related questionnaires consisting of items with constructed response formats (e.g. multiple choice, check list, rating scales) which are used as selection instruments for high volume jobs where an evaluation of experiences, education and other job related accomplishments is the appropriate examining strategy. This type of item format makes it possible to develop an optical scan form for automated processing and scoring of applicant responses. The cost and time estimates for processing optical scan forms are very similar to those required to process wiritten test answer sheets. The use of OSAAQ's makes it possible to process large numbers of applications in an economical manner and provides a standard measurement instrument that assures a high level of objectivity in the collection and evaluation of job related applicant information.

Most applicant questionnaires with objective scoring keys used in personnel selection have been based on a criterion related validity model (Van Rijin, 1980). Scoring keys are empirically derived for items which differentiate the job performance of incumbents. The research literature reports substantial support for the validity of questionnaire items from biographical information blanks and weighted application blanks based on empirically derived scoring keys (Owens, 1976). However, few of the studies cited in the literature provide any evidence for the job relatedness (i.e., item content samples important aspects of the job) of items used. It is not uncommon for background variables such as occupation/salary of father, place of residence, or age to be the type of items that are predictive of job success in many occupations. The likelihood of legal challenge, questions of test fairness and problems with public acceptance makes the use of non job related (from a content validity frame of reference) questionnaire items unacceptable in public sector selection programs regardless of their empirical validity. (Pace and Schoenfeldt, 1977).

Traditional T & E ratings are usually based on a content validity strategy and contain questions and scoring procedures which are closely related to the job studied. Typically, applicants provide narrative responses to questions on an application or supplemental form describing their experience, education and achievements relevent to the job. These narrative responses are reviewed by raters (preferrably trained subject matter experts) who assign points based on scoring protocals usually anchored with benchmarks describing several levels of experience, education and achievements. Requiring applicants to provide narrative responses to open ended questions and then utilizing rater judgements to evaluate these responses limits the objectivity of traditional T & E rating procedures (Levine and Flory, 1975).

The OSAAQ exam development methodology combines features from biographical information blanks and traditiona T & E ratings. Item formats consist of constructed response questions that result in a standard measurement instrument with an objective scoring key. The content of times is clearly job related and based on a content validation strategy. The development of OSAAQ items based on content validity strategy requires that specific procedures be followed to provide evidence for the job relatedness, clarity and appropriateness of items used to rank applicants for a particular occupation. Although these procedures usually require more effort than the development of traditional T & E ratings, they do not require the time and resources of an empirical validity study. When the volume of applications is high the "front end" work required to develop OSSAQ's is more than offset by savings in application processing and scoring.

Application

The OSSAAQ methodology is based on a content validity approach to exam development. Therefore, the application of this exam development methodology is most appropriate for situations where specific experience, education and accomplishments are typical of qualified applicants and are clearly related to important aspects of the job. For purely trainee level jobs where specific experience and education is not typical of applicants nor related to successful job performance, the OSAAQ methodology may not provide sufficient evidence to satisfy content validity requirements.

Studies are planned to investigate the feasibility of using OSAAQ's for entry level trainee jobs based on content validity. It may be possible to use these procedures on an interim basis pending the development of additional validity evidence.

Unassembled exams based on OSAAQ methodology are currently being developed for occupations which typically have large numbers of applicants for job vacancies. These occupations are also characterized by applicant populations that have specific experience, education or accomplishments relevant to important aspects of the work performed. One group of these occupations is high volume journeymen (full performance) level jobs with specific experience requirements which have a high degree of similarity in the work performed by incumbents. An example of this type of occupation for which an OSAAQ has been developed is Clerk Typist (GS-4) positions (Lyons, 1980). These are full performance level clerical positions which require applicants to have clerical experience. An occupational supplement with OSAAQ items was developed to measure applicants clerical work experience relevant to important aspects of the job. Try out of the OSAAQ and pilot testing of the exam procedures were completed before implementing the exam on a nationwide basis for selecting applicants for Federal clerical jobs.

A second group of occupations appropriate for the OSAAQ methodology is high volume entry level professional jobs which have specific education or experience requirements. These jobs usually require applicants to have some minimum amount of academic prepartation in the professional field in addition to related experience or college education. These basic requirements assure that applicants have common job related experiences or education relevant to important aspects of the work in the professional field. This makes if feasible to develop OSAAQ exams measuring job related experiences, education and accomplishments relevant to the applicant population and important to job success in the profession.

Exam development projects either completed or underway in this job category are Nurse, Forester, Biologist and Accountant/Auditor. For each of these occupations the format of OSAAQ items and the scoring procedures have been tailored to the grade levels covered (e.g. GS-4/9 for Nurse and GS5/7 for Forester), subspecialties identified and particular characteristics of the profession. The basic OSAAQ methodology is appropriate to all jobs in this category but the resulting item formats and scoring procedures are likely to vary from one occupation to another depending on the type of experience, education or accomplishments sampled and the way in which they contribute to job success.

A third group of occupations for which the OSAAQ methodology is appropriate are jobs in which applicants are likely to have experience, education, and accomplishments directly related to important aspects of the work but not required of applicants to be eligible for employment. An example of this type of job is entry level Computer Specialist. At the entry level there are no specific experience or education requirements for this job. However, many applicants have some computer related training or experience. From a technical point of view it is appropriate to use a content validity strategy for crediting job related past experience, education, and accomplishments in evaluating applicants for employment. As a practical matter it makes sense to consider highly job related experience and education common to many applicants in the examining plan for the occupation. For the Computer Specialist occupation OSAAQ items have been developed to obtain information about an applicant's education and experience for entry level positions. The OSAAQ can be used in conjuction with a written test of computer aptitude to rank applicants for entry level computer specialist positions. The combination of an OSAAQ with a written test results in a multiple assessment strategy which has the potential for greater validity than a single selection instruments.

Like other procedures based on a content validity strategy the application for the OSAAQ methodology to entry level trainee jobs is uncertain. It is difficult to identify past experience, education or accomplishments that can be considered a relevant sample of most trainee type jobs. Applicants for these jobs frequently do not have a common set of job related background experience. This limits the possibility of developing OSAAQ items which are appropriate to the applicant population and that sample relevant aspects of work in the occupation. At present OSAAQ development is limited to high volume occupations with applicant populations which have common background experiences that are directly related to the occupation studied.

Basic Requirements of the Model

The specific procedural steps for developing OSAAQs are likely to vary depending on the nature of the occupation studied, the characteristics of the applicant population and the availability of SME's and applicants during item development, try out and pilot testing. Although no one set of "cookbook" procedures is appropriate for OSAAQ development, the study design for each OSAAQ development project should have the following characteristics.

Behaviorally Based Job Analysis. Since the OSAAQ methodology is based on a content validity approach it requires a job analysis procedure that summarizes the worked performed in behavioral terms. A work behavior/task inventory is usually the most appropriate way of summarizing the important elements of a job. However, any procedure for describing the job in terms of observable behaviors important to job success is appropriate.

Job Related Items. The items developed for each OSAAQ must clearly relate to important aspects of the job. Items usually sample education, work experience, and other accomplishments which are directly related to important work behaviors required for success on the job. The closer the correspondence between the content of OSAAQ items and the work performed on the job the greater the support for content validity.

Objective Item Format. One of the main features in the OSAAQ methodology is the requirement that all items be of the constructed response type (e.g. multiple choice, check list, rating scale). An objective item format assures a standard measurement instrument which requires all applicants to respond to the same set of questions (see attachment 1 for an example). Open ended questions found on typical application or supplemental forms are not appropriate for the OSAAQ methodology because of the lack of standardization in the exam instrument and the inherent subjectivity present in scoring narrative responses.

Item Try Out. Items are assembled into an OSAAQ format and administered to job applicants usually as an experimental form along with the existing exam. Try Out of OSAAQ items provides information on:

- item relevance - the response rate to each item category is tabulated to determine the relevance of an item for the applicant population.
- item clarity - the type and number of errors made in responding to items are tabulated to check the adequacy of instructions and item format.
- item variability - a frequency count is made of responses to each item category to determine those items with adequate variability to differentiate among applicants.

Previous experience in developing OSAAQ exams has demonstrated the importance of this process in selecting appropriate items and developing scoring procedures.

SME Item Ratings. Subject Matter Experts (SME) are used to verify the job relatedness of OSAAQ items and to rate their value in differentiating between better and lesser qualified applicants. Scales for evaluating items are developed and independent ratings are obtained from SME's. These ratings along with other evidence from the job analysis and background information are used to establish item values for ranking purposes.

Objective Scoring Keys. Item selection and scoring of response alternatives are based on independent SME ratings and normative data collected during the try out with applicants. An objective scoring key is prepared for items that are consistently judged by SME's as differentiating among applicants on important job related factors and that have adequate variability based on try out results with applicants.

Pilot Studies. When time and resources permit it is desirable to conduct a pilot study of the OSAAQ exam before full scale implementation. Results of a pilot study are useful in refining the OSAAQ items and scoring key. The likely benefit from a pilot study is closely related to the results obtained during item try out. When significant changes in instructions or item format are suggested by the try out results further testing is usually required before implementing an OSAAQ exam.

Placement Follow-up Studies. The content validity approach used to develop OSAAQ exams is sufficient for documenting job relatedness requirements. However, past research on unassembled exam (Johnson, et. al. 1980; Schmidt et. al. 1979) provides scant evidence for the empirical validity of these procedures. Given the lack of empirical support for T & E ratings it is highly desirable to conduct placement follow-up studies to determine which item types and scoring procedures should be used to maximize predictive validity. Results from placement follow-up studies will provide evidence on the predictive validity of current exams and indicate changes required to improve the validity of OSAAQ exams. Pilot testing and placement follow-up studies are the only parts of the OSAAQ exam development process that are not required for establishing job relatedness based on content validity. However, these two parts of OSAAQ development should be carried out whenever feasible to improve the quality of the exam and establish its validity in predicting job success.

General Procedures

Although one set of procedures cannot be prescribed to fit all OSAAQ development projects, results from studies completed or in progress suggest a general approach for constructing OSSAQ exams which is practical to implement (reasonable time/resource requirements) and technically sound (job relatedness based on content validity). The procedural steps described in this section are intended as a guide to OSAAQ development. Exam development plans prepared for individual occupations will likely require some modification in the general procedures to assure that the basic requirements of the OSAAQ model have been met.

1. Review of Existing Occupational Information. For many Federal occupation there are a variety of informational sources about the work performed and about applicant characteristics important for successful job performance. Most of the information is likely to be found in unpublished reports, studies, manuals, or other institutional documents. Some examples are; OPM classification/qualification studies, agency position descriptions, special purpose job analysis studies, career surveys, certification requirements, training programs and completed application forms. A literature review of related journal articles books, and manuscripts is also valuable for identifing relevant published material. Besides collecting background information for the study, the review process provides the exam development specialist with occupational knowledge which is particularly important in planning and conducting a content validity study.

2. Prepare Work Behavior and OSAAQ Items. Background information collected about the occupation is used to develop a tentative draft of work behavior and OSAAQ items. Work behavior items consist of simple task statements describing one aspect of the work performed. The list of work behavior items should cover all important aspects of the work identified from the background information. Position descriptions, classification standards and special purpose job analysis studies are the typical sources of information for developing work behavior lists. Common sources of background information for developing OSAAQ items are completed application forms, training/education programs and certification requirements. These items consist of objective question with constructed

response alternatives covering specific work experience, education and other accomplishments directly related to work performed on the job. The content of OSAAQ items should be as observable and verifiable as possible. The items should sample experiences that are likely to be common to the majority of applicants. All work behavior and OSAAQ items that appear reasonable based on the background information reviewed should be listed. At this stage of the process the items should be exhaustive and in an objective format.

3. Panel Review by Subject Matter Experts (SME). A panel of SME's representing agencies/organizations with the largest number of selections is assembled to revise work behavior and OSAAQ items. The panel should consist of first line supervisors and/or work leaders of incumbents in the occupation studied. SME's should have a thorough knowledge of the job being studied and have direct contact with workers in the occupation. Panel members review the work behavior items for accuracy and completness. They also review OSAAQ items and recommend changes (revisions/additions) in the item pool. Additional items or revisions of existing items must be amenable to an objective format, deal with observable/verifiable job related experiences and be reasonable to expect from applicants.

4. Administer Work Behavior Inventory. Revised work behavior items are assembled into an inventory and administered to a representative sample of incumbents. Each work behavior item is rated by the incumbent on time spent and importance scales, as well as, indicating whether or not the item is performed independently, with guidance, or as a part of training.

5. SME Ratings of OSAAQ Items. An occupational supplemental consisting of revised OSAAQ items is prepared for review by a representative sample of SME's. A sample size of approximately 30 first line supervisors and/or work leaders representing major users is usually adequate depending on the number of agencies involved and the similarity of positions covered by the occupation. Independent ratings on the value of each item for differentiating between better and lesser qualified applicants is obtained from the SME's.

6. Try Out of OSAAQ Items. The OSAAQ items are assembled into a supplemental form and administered to applicants on an experimental basis along with current examining procedures. A representative sample of 100 completed forms from the try out group is usually adequate for each occupation, if placement follow up studies are not to be carried out with the try out group. When placement follow up studies are planned for the try out group, all applicants for a given period of time will be given the form to complete. The duration of time required is function of the selection ratio and the rate of placements.

7. Finalize OSAAQ and Develop Scoring Key. In selecting items for the final version of the OSAAQ both try out results and SME ratings are considered. Data from the try out should provide evidence of item relevance, clarity and variability. The SME ratings should confirm the job relatedness of items and provide estimates of the value of each response category for differentiating among better and lesser qualified applicants. Once the items have been selected and final revisions made a scoring key is developed based on SME ratings of item values. The distribution of scores from the try out group is used to develop a transmutation table for converting raw scores to final ratings.

8. Conduct Pilot Study (Optional). It may be desirable to apply the procedures on a small scale (e.g. two or three area offices) for a limited period of time to test out the operational characteristics of the exam before nation-wide implementation. This step can be valuable for checking revisions based on the try out results and for determining the effects of the scoring key and transmutation tables on the variability of applicant's ratings. Unforseen

administrative and operational problems can also be detected and appropriate changes made to correct them. Pilot testing is not an essential procedural step particularly when try out data and SME ratings result in little or no revision in OSAAQ items and when the scoring key produces reasonable variability in applicant ratings.

9. Placement Follow-up Studies. Documentation of exam development procedures based on the OSAAQ model provides adequate evidence for job relatedness based on content validity. However, further information needs to be collected to determine if OSAAQ items with a scoring key developed by these procedures is predictive of job success. Depending on the size of the try out sample and selection ratio it may be possible to collect placement follow-up data from the try out group. Since OSAAQ items are administered on an experimental basis to the try out sample there is less likelyhood of a restriction in range than with placement follow-up studies using operational examining instruments. When it is not practical (sample size and selection ratio are small) to conduct follow-up studies on the try out group, the studies can be conducted on applicants selected for jobs after the OSAAQ exam is implemented. Supervisory rating scales developed for each occupation studied are a practical method for collecting job performance data on applicant and ensure a standard criterion measure across organizational units studied.

The procedures described in this paper provide the basis for developing OSASQ exams. The procedures are practical to implement and have adapted well to current exam development efforts. However, the procedures are still developing and it is expected that improvements will be made in applying the OSAAQ model to exam development projects for other occupations.

Summary

The OSAAQ model provides a methodology for developing unassembled exams with objective item formats and scoring procedures. This makes it possible to utilize automated application processing and scoring capabilities to minimize the time spent by staffing personnel in routine application rating and register maintenance activities. Objective questionnaire formats and scoring keys also assures that applicant responses are consistently evaluated. Thus eliminating unreliability associated with the narrative response styles of applicants and with the judgements of examiners that is typical with traditional unassembled exams.

There are two limitations to the application of OSAAQ methodology for exam development. The first is a practical concern. Due to the "front end" work required to develop, try out, program and implement automated OSAAQ exams the occupation selected should not be expected to have a small volume of applications and selections. The second concern is a technical issue. Because the exam is based on a content validity strategy OSAAQ items must be based on job related accomplishments that are common to the life experience of the typical applicant. This means that for purely trainer jobs where there are no objective job related life experiences typical of applicant the OSAAQ methodology will usually be insufficient for documenting validity. Empirical validity studies are usually required to demonstrate the validity of exam procedures for trainee jobs.

The OSAAQ methodology provides a practical way of developing job related objective measurement instruments for situations where unassembled exams are appropriate. However, placement follow up studies are needed to evaluate the validity of item types and scoring procedures in predicting successful job performance.

REFERENCE

Johnson, J.C., Guffrey, W.L., & Perry, R.A. When is a T & E Rating Valid? Paper presented at the annual meeting of the IPMA Assessment Council, Boston, June 1980

Levine, E.L., & Flory A. Evaluation of job applications: A Conceptual framework. Public Personnel Management, November-December, 378-385, 1975

Lyons, T.J. Self Report Questionnaire for Selecting Typists: A Content Validity Approach. Paper presented at the annual meeting of the IPMA Assessment Council Boston, 1980

Owens, W.A. Background Data. In Dunnette, M.D. (Ed.) Handbook of Industrial and Organizational Psychology. Chicago: Rand McNalley, 1976

Pace, L.A., & Schoenfeldt, L.F. Legal concerns in the use of Weighted applications. Personnel Psychology, 30, 159-166, 1977

Schmidt, F.L., Caplan, J.R., Bemis, S.E., Decruir, R. Dunn, L., & Antone, L. The Behavioral Consistency Method of Unassembled Examining (Technical Memorandum 79-21) Washington, D.C: U.S. Office of Personnel Management, Personnel Research and Development Center, November 1979.

Van Rijin, P. Biographical Questionnaires and Scored Application Blank in Personnel Selection. Washington, D.C.: U.S. Office of Personnel Management, Personnel Research and Development Center, December 1980

D [illegible] ACADEMIC BACKGROUND

[illegible] education of [illegible]

[illegible] diploma [illegible]

[illegible] GED [illegible]

(2) [illegible] or Commercial School

[illegible] months

[illegible] months (less than 20 classroom hours per week)

[illegible] (at least 20 classroom hours per week)

[illegible] months (at least 20 classroom hours per week)

(3) College

[illegible]

[illegible] 30 semester hours or 45 quarter hours

[illegible] semester hours or 45 to 89 quarter hours

[illegible] hours or 90 or more quarter hours

[illegible] or higher degree

E NUMBER OF BUSINESS COURSES

Indicate the number of high school, business or commercial school, or college courses you have successfully completed in the following subjects: Typing, Stenography/Shorthand, Machine Transcription, Secretarial Practices, Office Management, Business English, Business Correspondence, Accounting/Bookkeeping, or Business Math.

Number of Courses	HIGH SCHOOL	BUSINESS SCHOOL	COLLEGE
None			
1 [illegible]			
[illegible]			
[illegible]			

F TOTAL LENGTH OF CLERICAL EXPERIENCE

Blacken **one** oval below to indicate the total length of **all** of your qualifying clerical experience. Qualifying experience includes jobs such as typist, stenographer, secretary, file or mail clerk, accounting or bookkeeping clerk, general office clerk, and the like. It **does not** include jobs such as sales clerk, driver, waiter, cashier, or other jobs which did not include clerical duties as a substantial part of the job. Part-time experience must be shown in terms of the equivalent amount of full-time (40 hours per week) experience.

[illegible] months [illegible] years more than 6 years

All experience shown in this section and in section G must be supported by the descriptions of your work experience on your Personal Qualifications Statement (SF-171) or other application forms.

G CLERICAL EXPERIENCE IN VARIOUS JOBS

In this section indicate the amount of experience you have had in each of the various types of jobs listed below. For each qualifying clerical job shown on your Personal Qualifications Statement (SF-171), pick the job title which **best** describes that job and blacken one of the ovals beside the job title to show how long you worked in that type of job. If you had more than one job of a particular type, blacken the oval which shows the total length of all your work experience in that type of job. For example, if you worked for 8 months as a file clerk and then had another job as a file clerk for 9 months, you should blacken the "15 - 17 months" oval opposite "File Clerk."

Be sure to pro-rate any part-time jobs. If you worked 20 hours per week for 18 months, you should show 9 months for that job.

Do not enter one job under more than one job title, even if the job did involve a mixture of duties. Pick the title which is the best overall description of the job. Use "General Office Clerk" to describe a clerical job that does not fit any of the other job titles.

	1-2 months	3-5 months	6-8 months	9-11 months	12-14 months	15-17 months	18-23 months	2 Years or more
1 GENERAL OFFICE CLERK								
2 TYPIST IN A TYPING POOL								
3 TYPIST FOR A SPECIFIC INDIVIDUAL OR GROUP								
4 FILE CLERK								
5 SECRETARY								
6 BOOKKEEPING ACCOUNTING CLERK								
7 KEYPUNCH OPERATOR DATA TRANSCRIBER								
8 RECEPTIONIST								
9 STENOGRAPHER								
10. MAIL CLERK								

DO NOT MARK IN THIS SPACE — FOR OFFICE USE ONLY

0 1 2 3 4 5 6 7 8 9

T S

C

C

7

AN ATTEMPT TO VALIDATE A SHORT JOB EVALUATION QUESTIONNAIRE

Walter Mann*
Office of Personnel Research and Development
U. S. Office of Personnel Management
Washington, D. C. 20415

SUMMARY

A job analysis undertaken for the Personnel Office of the U. S. Virgin Islands (V.I.) used questionnaire data provided by clerical and administrative personnel to accomplish several objectives. One objective was to develop a quantitative procedure for assigning grades to classes of positions by means of a 15-item job evaluation questionnaire. A factor analysis of the 15 job evaluation scales resulted in three interpretable factors: Autonomy, Complexity, and Environmental Demands. Factor scores for the first two of these factors were effective predictors of grade (R = .65). This correlation is fairly high considering the degree of error variance in the grades of individual positions. The use of more reliable class data, as would be the case in practice, should produce a higher correlation. To test this hypothesis, predictor scores for positions from the preceding analysis were grouped by grade; median predictor scores correlated .92 with actual grade. Despite the promising results, additional research is needed to validate the questionnaire.

*The statement expressed in this paper are those of the author and do not necessarily reflect official policies or opinions. Dr. Patrick McAuley participated in the development of the Job Analysis Questionnaire.

BACKGROUND

In the mid 1970's the U.S. Virgin Islands (V.I.) requested assistance from the U.S. Civil Service Commission (CSC) to upgrade their Personnel Office. The author was sent there on a one-year mobility assignment and initiated a variety of projects. One of these projects resulted in a multi-purpose job analysis of clerical and administrative positions. The intention was to provide V.I. management with a data base suitable for making selection, classification, and job evaluation decisions. A recently published report of that project (Mann, 1980) is oriented towards selection. The present paper will focus on job evaluation. The present paper will also describe the accomodations that had to be made because the job evaluation study was done within the context of a selection-oriented job analysis.

The V.I. study used the job evaluation factors and their definitions that had been developed for Federal, white-collar non-supervisory jobs from GS-1 through GS-15 (Anderson & Corts, 1973). These job evaluation factors and their definitions were based on a review of the literature and actual practice. They were reviewed by employee unions and professional organizations. A breakout of job factors into subfactors and definitions of degree levels for both was carried out by CSC. The five factors, with subfactors in parentheses, were: Knowledge Required by the Job, Responsibility (Supervisory Controls; Guidelines), Difficulty (Complexity; Scope and Effect), Personal Relationships (Personal Contact; Purpose), and Environmental Demands (Physical Requirements; Work Environment).

The Anderson and Corts questionnaire has one scale for each subfactor. A number of their scale points have definitions that appear to include multiple elements. In the present study only one element per scale was used, thereby making it easier for an untrained rater to comprehend and also providing a check on the internal consistency of each subfactor. Another change from the Anderson and Corts questionnaire was the deletion of content inappropriate to clerical and administrative personnel; e.g., Personal Contacts was dropped because it was judged to cover essentially the same domain as Purpose of Personal Contacts.

METHODOLOGY

The Questionnaire

The Job Analysis Questionnaire (JAQ) was developed specifically for this study. It contains 24 pages divided into five major parts: background information (15 items), knowledges (31), skills and abilities (104), job evaluations scales (15), and activities (99). The original form of the JAQ was reviewed by two personnel psychologists to determine its appropriateness in terms of comprehensiveness of coverage and understandability by subjects. It was pretested in two agencies: the Department of Health and the Personnel Office. As part of the pretest, subjects were instructed to suggest additional items. On the basis of the pretest the JAQ underwent extensive revision, especially with regard to instructions.

The JAQ is in Appendix A of Mann (1980). Only the parts pertinent to this report will be described here. The background section contains questions used to identify the sample or interpret the results, e.g., agency, length of experience, level of education, year of birth, sex, and class title. The 15 job evaluation scales were designed to measure 8 subfactors: Supervisory Controls (3 scales), Complexity (3), Guidelines (2), Scopes and Effect (2), Knowledge and Skill Required by the Job (2), Physical Requirements (1), Work Environment (1), and Purpose of Personal Contacts (1). Each scale had from three to five statements. Incumbents checked the one statement in each scale that best described their respective positions.

Procedure

The heads of all V.I. government agencies were sent letters requesting their participation in the study and asking them to name a project coordinator. Coordinators were asked to determine the number of full-time clerical and administrative personnel, GS-1 through GS-24, in their agencies. Sufficient JAQs were printed and delivered to each coordinator, along with instructions for administering. If so desired the JAQ could be completed anonymously.

The popluation was limited to full-time clerical and administrative employees in General Schedule grades 1 through 24 who had been in their jobs more than three months, and who worked in one of the 11 agencies that participated in the study. GS-24 was selected as the top grade because beyond this, to the limit of GS-30, are higher level professionals and managers. The General Schedule was chosen because it excluded those not of interest in the present study, such as blue-collar workers, police officers, firefighters, nurses, doctors, and teachers.

Sample

The sample consisted of 160 clerical and administrative employees in 78 classes. The individuals in the sample had been with the government anywhere from 3 months to 31 years (mean = 8 1/2 years), ranged from 18 to 60 years of age (mean = 36), had been in class from 3 months to 20 years (median = 4 years), ranged from GS-1 to 24 (median = 15), and had supervisors who ranged from GS-15 to 30 (median = 24). In this sample 77% were female, 96% had a high school diploma, and 13% had completed at least four years of college. In addition, 22% of the incumbents considered themselves journeymen, 8% master craftsmen, 18% lead workers, and 25% superviors.

Accurate population data were not available. Therefore, a comparison of sample and population data was not possible. The sample did not appear to be unrepresentative of clerical and administrative personnel in VI.

RESULTS AND DISCUSSION

Job Evaluation Dimensions

An attempt was made to reduce the 15 job evaluation to a few meaningful factors. Anderson and Corts (1973) hypothesized five such factors--Knowledge Required by the Job, Responsibility, Difficulty, Personal Relationships, and Environmental Demands--but no data to either confirm or deny. While the present study does not support the hypothesis of five independent dimensions, it does suggest a reasonable alternative.

The scores on the 15 job evaluation scales were intercorrelated (see Table 1). From this intercorrelation matrix, three principal component factors were extracted and rotated orthogonally. These three factors accounted for 55% of the common variance. The first factor, accounting for 27% of the common variance, had nine scales that loaded .5 or higher: Supervisory Controls (3), Guidelines (2), Scope and Effect (2), Purpose of Personal Contacts, and Knowledge Required by the Job (see Table 2). This factor was called Autonomy. On this first factor collasped the following Anderson and Corts factors: Responsibility, Knowledge Required by the Job, Personal Contacts, and half (Scope and Effect subfactor) of the Difficulty factor.

TABLE 1

Statistically Significant Intercorrelations of 15 Job Evaluation Scales

	Supervisory Controls			Guidelines		Scope		Personal Contacts	Knowledge	Skill	Complexity			Phys. Reqs.
	1	2	3	1	2	1	2				1	2	3	
Supervisory Controls 1														
" " 2	48													
" " 3	51	45												
Guidelines 1	47	45	24											
" 2	37	61	35	46										
Scope 1	45	49	39	41	40									
" 2	39	37	23	29	38	49								
Personal Contacts	32	37	37	30	25	42	35							
Knowledge Required	36	36	35	36	24	45	38	38						
Skill Required									31					
Complexity 1	35	50	37	28	33	53	37	33	38					
" 2	29	25	22	33	24	48	29	26	25	51				
" 3	36	37	41	38	27	48	38	34	36	58				63
Physical Requirements						26		21		-31				
Work Environment						31		28			24	22	22	38

Note: Decimal points have been omitted. To be significant at the .01 level, a zero-order product moment correlation based on an N of 160 must be at least .21.

TABLE 2

Factor Loadings of 15 Job Evaluation Scales (N = 160)

Job Evaluation Scale	Factor Loading I	Factor Loading II	Factor Loading III
Supervisory Controls 1	69		
" " 2	78		
" " 3	58		
Guidelines 1	62		
" 2	69		
Scores and Effect 1	55	49	
" " 2	51	33	
Personal Contacts	50		32
Knowledge Required by the Job	59	30	
Skill Required by the Job	43		-57
Complexity 1	47	60	
" 2		85	
" 3	30	80	
Physical Requirements			81
Work Environment			65

Note: Loadings less than .30 deleted. Decimals points deleted.

The second factor accounted for 16% of the common variance. Having only the Complexity scales loading .6 or higher, it was called Complexity. In the Anderson and Corts hypothetical framework, Complexity is only a subfactor, one half of the Difficulty factor.

The third factor accounted for 12% of the common variance. It had three scales loading .5 or higher: Physical Requirements, Work Environment, and Skill Required by the Job (negative loading). The first two scales comprise the Anderson and Corts factor called Environmental Demands, and therefore was given that name. The Skill Required by the Job scale was part of the Anderson and Corts Knowledge Required by the Job factor, but had to be removed from that factor because for clerical and administrative personnel the two differ.

Prediction of Grade

Step-wise multiple regression was used to estimate the extent that grade could be predicted from the data at hand. Factor scores for the three job evaluation factors were computed from the factor loadings and standardized scores of the 15 scales. Only the first two sets of factor scores were accepted by the step-wise multiple regression. The first set of job evaluation factor scores (Autonomy) correlated .55 with grade; the second (Complexity), .36. Together they correlated .65 with grade (significant at the .001 level).

Mann (1980) found considerable error variance in the grades of individual V. I. positions; based on an analysis of time spent on work activities, 31% of the positions were found to be misclassified. A more reliable measure of grade would theoretically correlate higher with the job evaluation factor scores than .65. To test this hypothesis, predictor scores for positions were grouped by grade; median predictor scores correlated .92 with actual grade.

RECOMMENDATIONS

Despite the promising results, the job evaluation questionnaire cannot be recommended for operational use until additional research has been completed. The present study was limited by a small sample size. For those interested in using the questionnaire the following recommendations are made:

(1) Consider the use of supervisors as subjects; they would have been used in the present study if it had been possible.

(2) Scale the 15 items in terms of the organization in which they are to be operationally used. Edit as appropriate.

(3) Don't apply the questionnaire to supervisory positions unless one or more scales on supervising others are developed.

(4) Perform the factor and regression analyses with a fairly large sample; 1000 would not be too many.

REFERENCES

Anderson, C. H., & Corts, D. B. Development of a framework for a factor-ranking benchmark system of job evaluation. Washington, D. C.: U. S. Civil Service Commission, 1973.

Mann, W. G. A multipurpose analysis of clerical and administrative positions in the U. S. Virgin Islands. Washington, D. C.: U. S. Office of Personnel Management, 1980.

AD P001349

REDUCING ADVERSE IMPACT VIA A MEASURE OF APPLICANT DISADVANTAGEDNESS

Walter G. Mann*
Personnel Research and Development Center
U.S. Office of Personnel Management
Washington, D.C. 20415

SUMMARY

Recent attempts to reduce the adverse impact of examinations have focused on alternatives to written tests. The present report, on the other hand, demonstrates how the adverse impact of written tests can be reduced by correcting for the degree to which a job applicant had been educationally and/or economically disadvantaged or deprived.

A measure of disadvantagedness containing nine items was internally and externally validated with 11,931 applicants for a nationwide, examination. Internal validation was demonstrated by a factor analysis that yielded two factors. External validation was based on the relationships of scores on the two factors with other variables: minority status, test performance, and educational level; each assumed relationship was confirmed.

A composite measure of the two factor scores, called D, had a point biserial correlation of -.31 with passing the test and .54 with minority status. The adverse impact of the test was substantially reduced by partialling D out of test performance. It was also demonstrated that partialling D out of test performance would not necessarily reduce the validity of the test, and could actually improve it.

INTRODUCTION

It has proven difficult to design valid selection procedures that have no adverse impact. Some employers believe the solution is to avoid the use of written tests. This approach, however, will fail to the extent that the alternative selection device has adverse impact, lacks validity, or is not cost effective.

The approach adopted in the present paper is not to eliminate the entire test--but to statistically partial out that part of it that is undesirable. If whites as a group do better on a test because they have had more advantages than minorities, then it is a simple matter to reduce the adverse impact of the test by statistically controlling for disadvantagedness. A measure of disadvantagedness is, of course, needed to make this correction.

*The statements expressed in this papers are those of the author and do not necessarily reflect official policies or opinions. Assisting at various stages of the project were John Kraft, Lois Northrop, Henrietta McClure, Anthony Mento, and Robert McKenzie.

The present paper addresses such question as:

- Can a measure of disadvantagedness be developed that is appropriate for applicants for a particular examination? Such applicants would, of course, be more homogeneous than the population at large.
- Can the adverse impact of a specific written test be reduced when the educational and economic disadvantagedness of applicants is partialled out?
- Must a test that does not handicap disadvantaged applicants be less valid than its conventional counterpart?

Assumptions

It was assumed (e.g., Bachman, 1972; Tyler, 1965) that disadvantagedness while one is growing up is related to the following:

- father's education and job (negative relationship);
- mother's education and job (negative relationship);
- lack of familiarity with the English language;
- type of community one grew up in (e.g., fewer advantages in the big city than in the suburbs);
- ethnic group (minorities more disadvantaged than whites--but intragroup differences);
- test performance as an adult (negative relationship);
- highest level of education (negative relationship).

METHOD

Procedure

Three self-administering questionnaires—CSC Form 1289-A (Temp), CSC Form 1310 (Temp), and CSC Form 1203-N--were given to all applicants nationwide who took a specific written test during 1978. Completed forms were optically scanned and the data recorded on computer tape. The author was assigned to the project after all the data had been collected.

CSC Form 1289-A (Temp) has seven questions: name, date, social security number, examining office code, race, sex, and ethnicity. Non-Hispanic whites were the majority group; all other ethnic groups were combined to form the minority group.

CSC Form 1310 (Temp) has 16 questions: name, date, year of birth, social security number, examining office code, type of mathematics courses taken, questions about conditions while they were growing up (type of community, whether English was the primary language spoken in home or school, and the perception of economic and educational advantage in the home), the most important occupation and the highest level of education of parents (or guardians), whether they or members of their family are currently employed by the Federal government, and whether they claim veteran preference.

CSC Form 1203-N has 10 questions: name, date, social security number, data available to start work, occupational specialty (applicants for another job took the same written test), regional preference, Spanish language ability, educational level, length of qualifying experience, and previous experience.

All the data generated by the first two forms were available to this investigator, with the exception of type of math courses taken. Data from the last form were available only for those applicants who passed the written test.

The written test consists of 65 items of the following types: 30 vocabulary, 15 English usage, 10 judgment, and 10 logical order of events. An applicant's score is the number of right answers; 35 is passing. Scores between 35 and 65 are transmuted on a point-by-point basis to ratings.

Sample

Data were obtained from 11,931 applicants. It was impossible to determine what percent this was of the total number of applicants; i.e., it was impossible to determine the return rate. However, Northrop (Note 1) found a 98% return rate for another nationwide examination.

The mean year of birth was 1951.8 (S.D. = 4.2). Females accounted for 9.2% of the sample. Included in the sample were 6821 whites, 3699 Hispanics, and 963 blacks.

Scaling the Items Used to Measure Disadvantagedness

The nine items used to measure disadvantagedness are from CSC Form 1310 (Temp) Nov. 1977. Although this form was developed by personnel psychologists with the intention of measuring socio-economic level, the items appear to be compatible with the measurement of disadvantagedness; it was only necessary to score them in the opposite direction from that originally intended. In fact, two of the items (concerning perceptions of economic and educational advantage) appear to be more closely related to disadvantagedness than to socio-economic level.

Scale values for these items were computed on half the sample (S1). Each alternative was given a value from 0 to 13 (least to greatest disadvantagedness, based on the judgment of the researcher).

A simple sum score (called initial D) of the nine items was computed for each individual in S1. The mean initial D for all individuals who answered an item in any particular way was computed. For example, the mean initial D of all those applicants who indicated they grew up on a farm was computed; likewise, the mean initial D for those who grew up in a small town; etc. For each item, the alternative with the lowest mean initial D was arbitrarily set to 0; and the alternative with the largest mean initial D, 20. Alternatives with intermediate initial D's were assigned whole numbers between 0 and 20 depending on their relative distances between the high and low points.

VALIDATION OF THE MEASURE OF DISADVANTAGEDNESS

Internal Validation of Disadvantagedness Items

Factor analysis was applied to the second half of the sample (S2) to develop evidence concerning the internal validity of the nine disadvantagedness items. Response to the items were assigned the scale values generated on S1. An iterative factor analysis of the intercorrelations between the items produced two significant roots. The program iterated until the number of significant roots was decided and the diagonal estimates stabilized. Varimax rotation was used to simplify the factor structure. The first factor accounted for 29% of the variance; the second factor, 11%.

Seven of the nine items, the highest of which were mother's education (.77) and father's education (.76), loaded at least .42 on the first factor (see Table 1). This factor was named General Disadvantagedness. Type of community (.98) was the only item that loaded higher than .08 on the second factor. This factor was named Community Disadvantagedness.

TABLE 1

Loadings of Two Factor Extracted from Intercorrelations Between Pairs of Disadvantagedness Items (N = 5965)

Questionnaire Item	Factor I	Factor II
Mother's Education	.77	.06
Father's Education	.76	.08
Father's Job	.69	.06
Spoken Language	.52	.08
Educational Advantage	.51	.05
Mother's Job	.51	.00
Economic Advantage	.42	.06
School Language	.16	.03
Type of Community	.10	.98

External Validation

An external check on whether a factor is measuring disadvantagedness is to determine its relationship with other variables. Factor scores based on standardized scores and factor loadings were correlated with test performance, minority status, level of education, sex and year of birth. From the correlations in Table 2, it is clear that high Factor I scores were associated with low test performance, being a member of a minority group, low educational level, and being old.

TABLE 2

Correlations of Two Disadvantagedness Factor Scores With Selected Dichotomized Variables

Selected Variable	Factor I	Factor II	N
Test Performance (pass = 1, fail =0)	-.30*	-.09*	5965
Minority Status (minority = 1, white = 0)	.52*	.12*	5877
Educational Level	-.11*	-.05*	2985
Sex (male = 1, female = 2)	.00	.00	5900
Year of Birth	-.15*	-.05*	5948

*$p < .01$

Factor II scores had the same correlational profile as Factor I scores, the only difference being the magnitude of the correlations. Both factors therefore exhibited the correlational profile that had been assumed to exist for disadvantagedness. The two sets of factor scores correlated -.022. Neither set of factor scores had a sex bias; however, both demonstrated a bias towards younger applicants. An overall measure of disadvantagedness, based on the two sets of factor scores was effected by a conventional multiple regression paradigm utilizing minority status as the dependent variable; the weights came from S1. The combination of the two factor scores will be referred to as D. D correlated -.31 with test performance and .54 with minority status.

ADVERSE IMPACT AND VALIDITY OF THE WRITTEN TEST

Use of D to Reduce Adverse Impact

The correlation of test performance with minority status is an easily computed, albeit crude, index of the adverse impact of a test. The four-fold point correlation between test performance (pass/fail) and minority status (minority/majority) was -.419 (significant at the .01 level).

D was partialled out of the test scores, producing in effect D-less test scores. The adverse impact of these D-less test scores was estimated by correlating them with minority status (r_p = -.335). Thus, controlling for D reduced the adverse impact of passing the test substantially--but hardly completely.

Neither the conventional test nor the D-less test demonstrated a significant sex bias. In fact, the correlation of test performance with sex was virtually zero.

Estimated Validity of D-Less Test Scores

Although the criterion-related validity of D-less test scores has not yet been determined, estimates of it can be generated based on assumptions about correlations of a criterion measure with (1) the conventional test and (2) D. It was assumed that the criterion correlation for the conventional test was between .2 and .4, and for D between .0 and -.2. Partialling D out of the conventional test scores and correlating the result with the criterion, produces the validity coefficients in Table 3. As the validity of the conventional test increases, so does the validity of the D-less test. In no case does the D-less test lose all its validity, not even when the conventional test validity is .20 and D correlates -.20 with the criterion. When D correlates .00 with the criterion, the validity of the D-less test actually exceeds the validity of the conventional test.

TABLE 3

Validity Coefficients of D-Less Test as a Function of Conventional Test Validity and Correlation of D with the Criterion

	Possible Validity of Conventional Test		
Possible Correlation of D with Criterion	.20	.30	.40
.00	.22	.33	.45
-.10	.19	.30	.41
-.20	.15	.26	.35

Note. Validity Coefficients are product moment correlations.

DISCUSSION AND CONCLUSIONS

The measurement of disadvantagedness in the present study was based on nine items. Although these items appeared to be effective indicators of disadvantagedness, they have limitations from a psychological point of view in that they are global or outputs, rather than processes or behaviors. For example, the type of community one grew up in is a very crude overall indicator of the presence of: libraries, museums, role models, reinforcement of intellectual behavior, and positive peer attitudes toward education. A better measure of disadvantagedness would rely on specific, psychologically rich indicators.

Because of the limitations of the items used to measure D and because of the limited sample (applicants for a single examination during 1978), generalization to other situations is dangerous. Future studies of other classes could provide the type of information needed to derive generalizations about the measurement of disadvantagedness and its relationship with constructs in the occupational setting.

The measure of D was based on research items. Before a measure of D could be used operationally, the items would have to be reprinted as part of the examination. The possibility that this could affect the candor of the applicants should be studied.

If D has low or negative validity the D-less test will have approximately the same validity as the conventional test.

REFERENCE NOTE

1. Northrop, Lois. Pilot study No. 1 for the collection of race, sex, and ethnic origin data in Federal testing. Washington, D.C.: U.S. Office of Personnel Management, 1977.

REFERENCES

Bachman, M. E. Patterns of mental abilities: ethnic, socioeconomic and sex differences. American Educational Research Journal, 1972, 9, 1-12.

Tyler, L. E. The psychology of human differences (3rd ed.) New York: Appleton - Century crafts, 1965.

Mathews, John J. & Roach, Bennie, W., Air Force Human Resources Laboratory, Brooks Air Force Base, Texas. (Withdrawn)

Development and Norming of Reading Tests for Air Force Use

This paper reviews development and norming of reading tests targeted at Air Force personnel. A number of different reading tests are in current Air Force use. These vary in intended target population from grades 4-6 (Gates-MacGinitie) to high school through college (Nelson-Denny). Such divergent level tests have yielded different normative scores for the same individuals (Mathews, Valentine, and Sellman, 1978). The purpose of reading tests developed by AFHRL is to provide standard tests targeted at and normed on Air Force personnel. Also, costs for test materials should be substantially less. Two equivalent Air Force Reading Abilities Test (AFRAT 2a and 2b) Forms have been developed for remedial reading programs. A broader level test (AFRAT 1) was used to obtain normative and validity data on reading items for 5,000 enlistees. Data reported include item and test statistics, relationships and test and demographic variables, and calibration with other reading tests.

DEVELOPMENT AND CALIBRATION OF ENLISTMENT SCREENING TEST (EST) FORMS 81a AND 81b

John J. Mathews and Dr. Malcolm James Ree
Air Force Human Resources Laboratory

I. INTRODUCTION

The Enlistment Screening Test (EST) was developed to reduce enlistment processing costs for transportation and boarding. By administering this test at local recruiting stations, those applicants who would most likely meet service mental qualification standards could be identified and sent to centralized testing stations.

Previous EST Forms 5 and 6 (Jensen & Valentine, 1976) were designed to predict Air Force qualification on the Armed Forces Qualification Test (AFQT) portion of the Armed Services Vocational Aptitude Battery (ASVAB). These ESTs became obsolete with the implementation of ASVAB Forms 8, 9, and 10 which do not have Space Perception items as did prior forms.

The objective of this study was to develop and norm two parallel EST forms for use by military recruiters in predicting applicant success on ASVAB 8, 9, and 10 selection composites. The new ESTs were designed in accordance with specifications which would make them appropriate for use by all armed services.

II. METHOD

Subjects

Service applicants were sequentially administered one of three experimental test booklets during March and April 1981 at geographically dispersed samples of recruiting stations representing each service. A total of 1,882 subjects had sufficient data forwarded in time to be included in the various analyses. The following subsamples were formed for specific analyses:

Sample 1. Subjects given booklet AX (N = 527) and for whom answer sheets were received by 24 April 1981.

Sample 2. Subjects given booklet BY (N = 486) and for whom answer sheets were received by 24 April 1981.

Sample 3. Subjects given booklet CZ (N = 457) and for whom answer sheets were received by 24 April 1981.

Sample 4. Subjects given booklet BY (N = 898) and for whom answer sheets were received by 14 May 1981. Sample 2 is a subset of Sample 4.

Instruments

In order to develop two parallel forms of an EST, each of which would require no more than 45 minutes to administer, three experimental booklets (AX, BY, and CZ) containing 60 items each and requiring 1 hour of administration time were assembled. Eight items were common to all booklets as a check on the comparability of samples given the different booklets. Each booklet contains 28 Word Knowledge (WK), 22 Arithmetic Reasoning (AR), and 10 Paragraph Comprehension (PC) items. These items were selected based on available data which indicated that they would discriminate best in the range of the 5th to 55th percentile for samples stratified on AFQT scores. The goal for EST items was to have about 40% with maximum discrimination around the 15th percentile, 40% with maximum discrimination around the 30th percentile, and 20% with maximum discrimination around the 45th percentile in the AFQT target population. This distribution maximized measurement at ability levels where most selection decisions are made. Items from the three experimental booklets were selected for two operational forms, each containing 40 to 50 items.

Data Editing

The following steps were taken to insure that only valid data were included in analyses:

1. Cases with no social security account number nor name were deleted.

2. Correct designation of booklet form was checked by scoring the test separately using the answer key for all three booklets. If an appreciably higher score would have been obtained based on the key for another form, then the booklet designation was changed. If the highest form score was below 8 (out of 60), indicating that the examinee was not trying on the test, the case was deleted.

3. Cases with no AFQT scores were deleted from all analyses except preliminary item analyses.

4. Cases with a standard error of estimate greater than $\pm$ 3.5 points from the predicted AFQT based on EST scores were deleted.

Less than 2% of the cases were eliminated from all analyses based on these editing procedures.

Data Analysis

Item statistics were generated to aid item selection for operational forms. Descriptive statistics including correlations and bivariate frequency distributions of EST with ASVAB selection composites were computed. The EST was calibrated to AFQT through equipercentile equating.

III. RESULTS AND DISCUSSION

Analyses on Stratified Samples

Samples 1 through 3 were rectilinearly stratified on AFQT percentile by random duplication of subjects so that an equal portion (10%) of the sample was represented in each AFQT decile. This procedure equated the samples on ability. Classical item analysis, including item difficulty and discrimination with AFQT as an external criterion, was then accomplished for each experimental form.

Items were selected to yield two parallel EST forms. All the items for one operational form (EST 81a) were chosen from the booklet (BY) which contained the greatest number of items in the appropriate difficulty and discrimination ranges. This procedure subsequently allowed direct generation of EST total score frequency distribution and bivariate statistics involving EST 81a and AFQT, since subjects in sample 4 were given all items which would be in that operational EST form.

Considerations in selecting items for the two forms included:

1. A significant positive correlation with AFQT.

2. Maximum discrimination in the desired ability range (5th to 55th percentile)

3. Content similar to AFQT (proportion of Verbal and Arithmetic Reasoning items).

4. Necessity of having the two forms parallel.

Two EST forms of 48 items each which met the desired specifications were constructed. A comparison of the content of these forms, designated EST 81a and 81b, and AFQT is given in Table 1. Because of administration difficulties, it was decided to exclude the speeded Numerical Operations (NO) item types from the EST. After deleting NO, AR items equally comprise 37.5% of AFQT and EST content. There are relatively more WK and fewer PC items in EST than in AFQT. However, these two verbal item types are highly intercorrelated. Because WK items require less time to complete, an abundance of these items saves testing time.

Difficulty (p) levels of the two forms based on stratified samples are indicated in Table 2. Item p's range from .57 to .89. Mean $\bar{p}$'s for the ESTs are virtually identical (.744 and .742), and the distributions are similar.

Biserial correlations (validity estimates) of items with AFQT percentiles ranged from .29 to .63, with about 75% between .45 and .59 (see Table 3).

Again, the forms appear quite comparable, with similar biserial r means (.476 and .496) and distributions. An internal consistency reliability (KR-20) of .93 was obtained for EST 81a. Since no subjects were administered all items in EST 81b, test statistics, including reliability, were not computed for this form.

Item Response Theory (IRT) item analytic indexes (Lord & Novick, 1968) were also computed for EST 81a based on the Birnbaum (1968) three-parameter logistic model. The three indexes are: a (item discrimination), b (item difficulty in Z metric), and c (probability of guessing) (see Ree, 1979, for a detailed description of these item parameters). Two types of analyses were completed. The first was based on EST score as an internal criterion, and the second was based on AFQT percentiles transformed into Z scores for the sample. Table 4 presents the mean a, b, and c values for these analyses on EST 81a.

As would be expected, the mean a ($\bar{a}$) was somewhat higher in the internal compared to external criterion analysis, although both $\bar{a}$'s were relatively high (1.3 and 1.1, respectively). Both $\bar{b}$'s were -.62, corresponding to a mean percentile of 27. The mean of the desired (targeted) distribution of b would be about -.67. The $\bar{c}$ was .20 based on the internal analysis. The classical probability of guessing for a four-choice item is .25. The $\bar{c}$ based on AFQT ability estimates was somewhat higher, .32.

Table 5 gives the mean and standard deviation (SD) of EST 81a based on sample 2 stratified on AFQT. Summary statistics for EST 81b would be quite comparable, since the available data indicate it is indeed parallel to EST 81a.

Intercorrelations are also shown in Table 5. The Pearson product-moment correlation (r) of EST 81a with AFQT percentile is .83. This indicates that about 69% of the variance in AFQT scores can be predicted from EST scores. The two EST subscales correlate with EST total to about the same degree as the like-named ASVAB composites correlate with AFQT. The r's of the Verbal (VE) and Arithmetic Reasoning (AR) scales with EST total are .95 and .88, respectively. The corresponding r's of ASVAB 8, 9, and 10 VE and AR with ASVAB 8, 9, and 10 AFQT are .93 and .89, respectively (Ree, Mathews, Mullins, & Massey, 1981).

Distributional Statistics

An equipercentile calibration was accomplished for the purpose of converting EST scores to equivalent AFQT percentiles. Table 6 gives the computed percentiles for sample 4 (N = 869) on EST 81a and AFQT and the AFQT percentile (from ASVAB conversion tables) which is equivalent to the computed percentile for each EST score.

The EST score distribution is negatively skewed, indicating potential for good discrimination at lower ability levels, which was one of the test construction goals. The median EST score is 40 (out of 48 points) and this is equivalent to an AFQT percentile score of 50.

The percentages of subjects within various AFQT score categories is shown for EST scores (grouped in intervals to increase Ns) in Table 7. The AFQT categories represent various service cutoff score boundaries.

Relationship of EST to General Composites

The correlation between the EST and the General-Technical (General for Air Force) composite for subjects given booklet BY in March 1981 (N = 270) was .86. This strong relationship was expected due to similarity in content. Since percentile equivalents for this composite are based on the same reference test used in norming AFQT and since it correlates highly with AFQT (r = .97), the equipercentile calibration (Table 6) would also apply to General (GT) percentiles.

IV. CONCLUSIONS

Two parallel forms of an EST have been developed and calibrated to the AFQT selection composite from ASVAB 8, 9, and 10. These tests appear to meet administrative and psychometric specifications adequately. All items in both forms correlate positively with total test score and AFQT scores and are in an appropriate range of difficulty (from average to very easy) for use in prescreening service applicants.

The EST should be a highly reliable instrument (internal consistency coefficient = .93). The test items appear to discriminate well throughout a range which includes major service selection cutoff points (AFQT percentiles 15 to 45). The two EST forms appear parallel based on highly similar distributions of item difficulty and criterion correlation values. EST scores predict AFQT percentiles quite well (r = .83). In addition, EST content is similar to that of AFQT.

Interpretation of EST scores is provided by an equipercentile calibration to AFQT (Table 6). This calibration also applies to prediction of GT percentile scores.

REFERENCES

Birnbaum, A. Some latent-trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

Jensen, H. E., & Valentine, L. D., Jr. *Development of the Enlistment Screening Test-EST Forms 5 and 6*. AFHRL-TR-76-42, AD-A033 303. Lackland AFB, TX: Air Force Human Resources Laboratory, Personnel Research Division, May 1976.

Ree, M. J., Mullins, C. J., Mathews, J. J., & Massey, R. H. *Armed Services Vocational Aptitude Battery: Item and factor analyses of forms 8, 9, and 10*. AFHRL-TR-81-XX. Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division, 1981, in press.

Lord, F., & Novick, M. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

Ree, M. J. Estimating item characteristic curves. *Applied Psychological Measurement*, 1979, *3*, 371-385.

Table 1. Content of AFQT and EST 81a - 81b

Item Types	Number of Items AFQT	EST
Arithmetic Reasoning (AR)	30	18
Word Knowledge (WK)	35	23
Paragraph Comprehension (PC)	15	7
(Verbal = WK + PC)	(50)	(30)
Numerical Operations (NO)	50*	-
Total # of items	130	48

*Speeded test with 50 items. NO raw scores are weighted .5 in the AFQT composite.

Table 2. Distributions of Item Difficulties for EST 81a and 81b

	81a (Booklet BY)		81b (Booklets AX & CZ)	
Difficulty (p)	N	%	N	%
.80 - .89	12	25	12	25
.70 - .79	23	48	19	40
.60 - .69	12	25	16	33
.00 - .59	1	2	1	2
Mean p	.744		.742	

Table 3. Distributions of EST Item Validities with AFQT

Validity (r_{bis})	81a (Booklet BY) N	81a (Booklet BY) %	81b (Booklets AX & CZ) N	81b (Booklets AX & CZ) %
.60 - .99	2	4	3	6
.45 - .59	35	73	35	73
.30 - .44	11	23	11	23
.00 - .29	1	2	0	0
Mean $r_{biserial}$*	.476		.496	

*Based on r to Z transformations.

Table 4. Means of IRT Item Parameters for EST 81a

	$\bar{a}$	$\bar{b}$	$\bar{c}$
Internal Criterion	1.33	-.62	.20
AFQT Criterion	1.10	-.62	.32

Table 5. Descriptive Statistics and Correlations for EST 81a in an AFQT-Stratified Sample (N = 445)

Scale	Mean	Standard Deviation	Intercorrelations VE	AR	EST	AFQT
Verbal Ability (VE)	22.7	7.4	1.00	.69	.95	.77
Arithmetic Reasoning (AR)	13.0	4.7		1.00	.88	.75
EST Total	35.7	11.1			1.00	.83
AFQT Percentile	49.5	28.4				1.00

Table 6. Equipercentile Calibration of EST to AFQT (N = 869)

EST Score	Cumulative % Below EST Score	AFQT Percentile	Cumulative % Below AFQT Score
1-11	1.4	4	0.9
12	1.7	5	1.5
13	2.3	6	2.6
14	2.8	6	2.6
15	3.2	8	3.3
16	3.6	8	3.3
17	4.6	10	4.5
18	5.4	12	5.5
19	6.2	13	6.2
20	6.9	14	7.4
21	7.4	14	7.4
22	8.5	15	8.6
23	9.8	15	8.6
24	10.6	16	10.9
25	11.5	18	12.1
26	12.7	19	12.8
27	14.4	20	14.8
28	16.0	22	16.6
29	17.5	23	17.7
30	19.9	25	20.6
31	22.0	28	22.9
32	24.2	30	25.5
33	27.4	33	28.8
34	30.6	36	32.5
35	34.3	40	36.0
36	38.0	42	38.4
37	41.1	44	41.4
38	44.9	48	45.3
39	47.6	49	47.8
40	50.5	50	50.5
41	56.5	54	56.8
42	60.4	59	62.6
43	64.8	63	66.9
44	70.0	68	72.0
45	75.5	74	77.6
46	82.9	80	85.4
47	90.0	87	92.8
48	96.0	95	98.0

Table 7. Distribution of EST Scores by AFQT Category

EST Score	AFQT Category 1-15 N	%	16-20 N	%	21-30 N	%	31-49 N	%	50-64 N	%	65 & > N	%	N
1-12	7	90.0	2	10.0	-	-	-	-	-	-	-	-	20
13-14	7	87.5	-	-	-	-	1	12.5	-	-	-	-	8
15-16	6	50.0	3	25.0	3	25.0	-	-	-	-	-	-	12
17-18	9	64 3	3	21.4	2	21.4	-	-	-	-	-	-	14
19-20	5	50.0	4	40.0	1	10.0	-	-	-	-	-	-	10
21-22	5	23.8	7	33.3	6	28.6	2	9.5	1	4.8	-	-	21
23-24	4	26.7	5	33.3	3	20.0	3	20.0	-	-	-	-	15
24-26	4	16.0	6	24.0	5	20.0	6	24.0	3	12.0	1	4.0	25
27-28	3	11.1	6	22.2	6	22.2	9	33.3	3	11.0	-	-	27
29-30	5	12.8	6	15.4	12	30.8	12	30.8	2	5.1	2	5.1	39
31-32	2	4.3	4	8.5	11	23.4	22	46.8	6	12.8	2	4.3	47
33-34	5	8.3	4	6.7	18	30.0	23	38.3	7	11.7	3	5.0	60
35-36	1	1.7	2	3.4	8	13.6	27	45.8	18	30.5	3	5.1	59
37-38	1	1.8	1	1.8	11	19.3	22	38.6	18	31.6	4	7.0	57
39-40	-	-	1	1.3	5	6.5	31	40.3	23.	29.9	17	22.1	77
41-42	-	-	-	-	-	-	22	30.6	24	33.3	26	36.1	72
43-44	-	-	-	-	2	2.2	10	10.8	34	36.6	47	50.5	93
45-46	-	-	-	-	-	-	3	2.4	21	16.7	102	81.0	126
47-48	-	-	-	-	-	-	-	-	6	6.9	81	93.1	87
N	75		54		93		193		166		288		869

7

Twenty Years Later: A Follow-up Of The Officer Evaluation Center

Pamela V. Mays, Ph.D.
Frederick Dyer, Ph.D.

U.S. Army Research Institute Field Unit
Ft. Benning, Georgia 31905

Abstract

After about 20 years, a follow-up was done of the Officer Evaluation Center (OEC), where 900 first and second lieutenants underwent 3 days of assessment of their technical, administrative and combat skills. From 15 exercises, over 2000 measures were taken of each officer. These observations were reduced through factor and other analyses to 341 variables which yielded 8 cross-situational factors.

During the current follow-up, the ability of 25 remaining summary variables to discriminate between the group of officers who left the Army after their initial 2-year commitment and the group remaining for a full career term was tested. A second discriminant analysis was performed among those who left active duty after 2 years. These were grouped according to rank (1st Lieutenant or Captain) at the time of completing their reserve commitments.

Both analyses yielded significant discriminant functions. In the former (active vs. discharged) analysis 65.38% of cases were correctly classified and in the latter 67.87%.

Background

In the late 1950's and early 1960's research was conducted by the U.S. Army Personnel Research Office to develop means of identifying officers with the aptitudes and characteristics to successfully meet the demands of different types of command responsibility. In essence, the research program centered around the development of the Differential Officer Battery (DOB). This battery included measures of information ranging from military tactics to the physical sciences, sports and the arts. Biographical reports and self-descriptive statements of interests and attitudes were also included. In the process of development and refinement, the battery was administered to 6500 active duty officers in 1958 and 1959 and about 4000 in 1961 and 1962 (Helme, Willemin and Grafton, 1971).

Suitable criterion measures were needed to validate this instrument. Ratings by peers and superiors were used as part of the validation effort. However, these were not totally satisfactory in that the DOB had been designed to differentially assess potential for combat, technical and administrative assignments. An officer's job rating was relevant only to his current assignment which could be representative of only one of the three categories.

It was decided that a series of situational tests would be administered to serve as additional validation criteria. These would allow assessment of each officer in each of the three areas and provide the added advantage of uniformity of tasks and standardization of observations.

For the purpose of administration of these situational tests, the Officer Evaluation Center (OEC) was established at Fort McClellan, Alabama on 1 March 1962. The first year of the center's operation was spent staffing, training assessors and finalizing procedures. The first officers who had taken the DOB were not tested until February of 1963. Final revisions were made based on this "shakedown" sample and for-the-record testing began in June of 1963.

In the process of refinement, all OEC exercises had been worked into a central scenario. This framework was that of a simulated Military Assistance Advisory Group (MAAG) Headquarters. New assessees were told to assume that they were "reporting for duty" at this MAAG Headquarters located in a friendly host nation. All tests then became a succession of assignments to be performed while temporarily awaiting reassignment to a field unit (Willemin, 1964).

Exercises were selected to provide reliable although not necessarily complete coverage of the technical, administrative and combat areas. All exercises had to meet certain conditions. They were required to be able to be performed without specialized training and experience, to be recognizable as representative military requirements, and to have militarily meaningful outcomes characteristic of good or poor performance.

Exercises were drafted with the assistance of subject-matter experts, field tested and then technically reviewed at the appropriate branch schools. They were designed to include measures of the following categories of behaviors: perceiving situational elements, judging future developments, analyzing problem elements, planning future action, organizing resources, deciding the course of immediate action, taking the initiative to act, communicating orders and information, training and directing subordinates, and persisting under stress (Willemin, 1964).

Each exercise was to be primarily representative of one of the three areas of interest. There were five exercises developed in each of the three areas. A summary of these is given as follows:

Combat Exercises:

1. March Order. Examinee plans a tactical road march and reacts to interruptions by senior and subordinate personnel.

2. Observation Post. Examinee directs fire onto visible targets. He must perceive terrain, enemy activity and targets; estimate range and communicate this information.

3. Security Mission. Examinee must anticipate enemy actions, quickly plan offensive and defensive actions and direct subordinates through face-to-face contact.

4. Roadblock. Examinee must apply basic tactical principles and communicate important information to others.

5. Route Reconnaissance Patrol. Examinee must cope with persistent obstructions to mission progress, respond to critical situational factors and withstand psychological stress under simulated prisoner-of-war conditions.

Technical Exercises:

1. Communications Exhibit. Examinee trouble shoots technical equipment and must use subordinates as effectively as possible.

2. Automotive Inspection. Examinee detects equipment deficiencies and recommends and/or performs corrective actions.

3. Road Damage and Radiation Survey. Examinee must organize teams, train subordinates, collect and communicate information and make plans under conditions of time pressure, obstacles, harassment and fatigue.

4. Airfield Layout. Examinee must use technical information to select an airfield site and compute the necessary length of a runway.

5. Weapons Assessment. Examinee reports on the characteristics of an enemy weapon from a technical intelligence point of view.

Administrative Exercises:

1. Improper Supply Records. Examinee analyzes supply records, writes a summary memorandum and (tactfully) communicates discrepancies .

2. Office Management. Examinee must organize administrative tasks and correct improper office procedures.

3. Production Analysis. Examinee analyzes production data, organizes unit for efficient operation, and communicates plans.

4. Site Selection. Examinee must use logistical judgment to interpret information and consider factors in site selection.

5. Highway Traffic Plan. Examinee must plan logistical support for a large scale tactical operation and respond to rapid political and military changes.

Each officer went through the exercises as an individual. The entire set required 3 days to administer. The combat setting was made as realistic as possible with 17 officers and 41 enlisted personnel playing the roles of United States, allied and enlisted personnel. The first day's exercises were carried out under time pressure but "peacetime" conditions. On the second day the examinee was awakened at 0230 after about four hours sleep and told that the host nation was at war. The remainder of the exercises were carried out under "emergency" conditions and increasing fatigue on the part of the examinee (Helme, Willemin and Grafton, 1971).

Method

Sample

The original sample of OEC participants was drawn from the pool of 4000 lieutenants who took the DOB between 1961 and 1964. Of these, about 900 attended the OEC after one or two years of active duty. Both first and second lieutenants were included as were graduates of the U.S. Military Academy and both Reserve and Regular Army graduates of Reserve Officer Training Corps (ROTC). The lieutenants represented 10 different combat arms, combat support and combat service support branches. Only about 737 of the original 900 participants are included in the data base. The remaining officers were members of the first thirty-odd groups used as a "shakedown sample" to refine measures and exercises (Helme, Willemin and Grafton, 1971).

The first step of the current research was to determine where these 737 men were in relation to their military careers and what data were available to indicate whether their performances at the OEC bore any relationship to their later degrees of military success.

Through the Army's locator service, we were able to find the names of 101 OEC participant officers still on active duty. At the time of follow-up sampling (1980) these included 1 Colonel, 86 Lieutenant Colonels, 11 Majors and 3 whose current ranks were indeterminate from information provided. The names of 412 additional OEC participants were found through computer search at the National Personnel Records Center (NPRC) in St. Louis, Missouri. The location of their records at NPRC indicated that these men had been discharged from all active and/or reserve military commitments.

The military history of the remaining 224 participants may be considered unknown. However, there is a third major repository of military records which is the Reserve Component Personnel and Administrative Center (RCPAC) in St. Louis, Missouri. This center houses records of individuals involved with Reserve Component (National Guard, etc.) units. It is not unlikely that many of the remaining OEC records could be found there, but we have as yet not been able to make suitable arrangement to obtain information from this center.

Information Gathered

It was quickly determined that a limited amount of information would be available for the "discharged" subsample located at NPRC. A much greater variety of information is available for the subsample of officers still on active duty which would necessitate a much more thorough process of development for a "criterion-of-success" score. Therefore, it was decided to obtain available information on the "discharged" group as a first step.

We found only certain useful forms to be contained as a rule in the majority of NPRC folders. These were:

Form DD 214--Report of Transfer or Discharge
Form USAAC 872--Discharge
Form 67-5, 67-6--US Army Officer Evaluation Report

Only those items were taken from these forms which might reasonably be considered indicative of military success. These were: 1) Number of years of active military service; 2) rank at the time of discharge from active duty; 3) rank at the time of discharge from the reserve component; 4) reason for discharge from the reserve component; 5) Officer Evaluation Report totals.

OEC Summary Variables

During the conduct of the assessment center more than 2000 observations and judgments were recorded on each assessee. These consisted of checklists of specific behaviors, scale ratings and quantitative summations of written products. Initially these items were analyzed by factor analyses conducted separately for each exercise. Intercorrelation and factor analysis of these scores yielded 342 scales or variables.

The number of variables was then reduced to 256 by elimination of those which were linear combinations of less complex ones and those on which 90% or more of the participants scored alike. Further factor analysis resulted in the identification of a set of 30 factors, all but two of which were specific to a single task.

To find cross task factors, "marker" variables were chosen for each factor. These were then combined with additional independent scales, refactored and rotated. A set of eight factors was identified and analysis using these 8 factors was then extended to the remaining variables (Helme, Willemin and Grafton, 1971).

Information remaining from the original set of OEC data consists of 25 summary variables. These scores represent 7 of the original 15 exercises (3 from administrative exercises and 2 each from the combat and technical areas). These summary scales were part of the 342 variables derived in the initial set of analyses. About half are shown as loading on the final 8 cross-situational factors derived from analyses. Few (about 5) are seen as markers or variables loading on the intermediate set of 30 variables. It is likely that many of them were omitted from this stage of analysis because they were linear combinations of simpler variables. A summary description of the variables is provided at Table 1.

Measures of leader characteristics resulting from Differential Officer Battery (DOB) development were correlated with OEC variables and factor scores. A number of significant correlations were found and differential prediction of the combat and technical-managerial leadership domains was shown (Helme, Willemin and Grafton, 1974).

TABLE I

SUMMARY DESCRIPTION OF REMAINING OEC VARIABLES

Area	Exercise	Variable	Loads on Cross-Task Factor
Administrative	Highway Traffic Plan	Factor Total	8-Technical Skills
		Attention to Data Requirements	1-Technical Managerial Leadership
	Office Management	Sequencing of Operations	[A]
		Retained Procedures	
	Site Selection	Factor Total	
Technical	Automotive Inspection	Factor Total	8-Technical Skills
		Identifying Information	5-Mission Persistence
	Airfield Layout	Sites Weighted Scale	
		Basic Geographical Considerations	
		Operational Hazards	
		Engineering Considerations	[B]
		Computational Accuracy	
		Utilization of Terrain Features	[C]
		Number of Sites Evaluated	7-Tactical Skills
		Thoroughness of Runway Report	
		Total Score	7-Tactical Skills
Combat	Security Mission	Firm Handling of Personnel	
		Effectiveness of Defense Plan	
		Total Score	2-Combat Leadership
	Roadblock	Attitude & Motivation	
		Tactical Control	
		Instruction of Men	3-Team Leadership
		Handling of Sniper	
		Confidence & Forcefulness	2-Combat Leadership
		Effectiveness in Establishing Abatis	3-Team Leadership

[A] Marker for intermediate factor 30-Commo & Staff

[B] Independent variable

[C] Marker for intermediate factor 23-Mission Accomplishment

Results and Discussion

It was determined that the best use of the existing data would be to determine how effectively the OEC variables could discriminate between the group of participants who chose to get out of the Army after their initial obligation and the group who decided to remain for a full career term. The decision(s) to remain in the Army is the fundamental criterion of a successful military career. It is the summary outcome of all the skills, motivations, experienced successes, etc. which allow one to choose and successfully complete a given life's work. Any set of variables potentially able to detect fine differences in level of success such as one-time ratings or awards should also be able to detect differences in this basic yet overriding criterion.

The group of 101 career officers for the analysis was self-defined. However, the discharged group required some further definition. Of 412 cases available, we were actually able to get data on 352. Of these by far the majority (237) fit the pattern of minimal 2 year active duty commitment and completion of the remainder of their obligation in some type of a reserve unit. (As previously mentioned, we have not been able to obtain data on anyone who may still be maintaining his reserve status.)

It was decided to use the homogeneous sample of 237 for the second group. An informal perusal of the records indicated that those having more or less than two years of active duty represented a much more ill-defined group. These included: officers killed in Vietnam, West Point Graduates leaving after their minimal 5-year commitment, medical discharges and a variety of unique cases.

A stepwise discriminant analysis was performed using the "2-year" and "20-year" career groups described and a significant discriminant function was found. The value of Wilks' lambda was .89 with a corresponding chi-square of 35.54 (d.f. = 7; $p < .001$). The canonical correlation was .318. However, neither of these statistics indicates a very high degree of separation between the groups.

Standardized function coefficients are shown at Table 2 for the 7 variables of the total 25 included in the function. These show the relative contributions of each variable to the function. By looking back to Table 1, one can determine the factors from the original analysis on which these variables loaded. It is interesting to note that while only 10 of the 25 Summary variables were reported as loading on the final factors of the original analysis, 5 of the 7 appearing in the current analysis came from these 10.

The cross-comparison also helps to lend interpretation to the function. To the extent that these variables are indicative of the original factors of combat and team leadership, tactical skills and mission persistence, the military careerists appear to be distinguished from the other group along a general "military leadership" dimension.

TABLE 2

Standardized Discriminant Function Coefficients and Group Means

First Discriminant Analysis (2 vs. 20 years)

Variable	Standardized Coefficient	Group Means 20-yr.Active	Group Means 2-yr.Active
Airfield Layout-No. of sites evaluated [1a - 2]	-.50	5.06	5.24
Automotive Inspect-Factor Total [1b - 2]	.24	9.34	8.16
Roadblock-Confidence & Forcefulness [1c]	.80	28.61	25.03
Security Mission-Total Score	.44	297.71	254.11
Airfield Layout-Comput. Accuracy	-.30	.55	.64
Roadblock-Instruction of Men [1d]	-.34	10.71	9.89
Site Selection-Factor Total	.23	10.12	9.35

Second Discriminant Analysis (Discharged as 1LT vs. CPT)

Variable	Standardized Coefficient	Group Means Captain	Group Means 1st Lieutenant
Airfield Layout-No. of sites evaluated [1a - 2]	-.38	5.18	5.33
Automotive Inspection-Factor Total [1b - 2]	.53	9.25	7.53
Roadblock-Attitude & Motivation	.57	29.48	27.20
Roadblock-Handling of Sniper	-.32	4.82	5.26
Roadblock-Tactical Control	-.38	3.12	3.34
Highway Traffic Plan-Attn. to Data [1e]	-.29	3.88	4.23
Airfield Layout-Util. Requirements of Terrain Features	.40	1.67	1.45

[1] These variables loaded on factors in the original analyses: a) Tactical Skills; b) Mission Persistence; c) Combat Leadership; d) Team leadership; e) Technical-Managerial Leadership.

[2] These variables included in both discriminant functions.

Following determination of the discriminant function, its ability to correctly classify cases was examined. Assuming the prior probability of group membership to be at the chance (50%-50%) level, 65.38% of cases were correctly classified using the discriminant function. This represents a 30.76% improvement over chance. Classification results are shown below:

	Predicted Career Group	
Actual Career Group	20-year	2-year
20-year	68 (67.3%)	33 (32.7%)
2-year	84 (35.4%)	153 (64.6%)

Lack of information and the considerable time span involved make it difficult to discuss those considerations that normally go with classification. For example, the large difference in group sizes would suggest improvement in overall classification through the use of prior probabilities of group membership other than chance. However, the most appropriate percentages to use were not readily available to us. They would be the statistical projections of officer retention of twenty years ago.

Those factors affecting tolerance for misclassification have also changed. The Selective Service System was still in effect in the early 1950's. Under that system the loss of a potentially successful officer through misclassification might have been much less costly than it is today.

Following the initial analysis, the 2-year sample of officers deciding not to remain in the Army was examined. This sample revealed a bimodal distribution along the dimension of rank at the time of discharge from the reserves. Of the 222 officers for whom records were available, 84 were discharged as Captains and 137 as First Lieutenants.

Assuming this to be an indicator of military success, a second stepwise discriminant analyses was performed using groups formed on the basis of rank at the time of reserve discharge. A significant discriminant function was found with Wilks' lambda = .90, chi-square = 21.58, d.f. = 7, p = .003 (canonical correlation = .308).

The accuracy of classification was checked and 67.87% of cases were correctly classified for an improvement of 35.74 % over chance. Classification results are shown below.

	Predicted Rank at Discharge	
Actual Rank at Discharge	1LT	CPT
1LT	93 (67.9%)	44 (32.1%)
CPT	27 (32.1%)	57 (67.9%)

Standardized discriminant function coefficients shown at Table 2 indicate the relative contributions of variables to the function. As with the first analysis, exactly 7 of the 25 variables are included in the function. However, except for the two variables marked, sets of variables belonging to the two separate analyses do not overlap. Also the variables in the latter analysis tend not to be the ones which loaded on factors in the original analysis. (The exceptions are the two overlapping variables mentioned and the "highway" variable.)

This would suggest that the dimension(s) separating the career-bound young officer from the one who will leave for civilian life may not be entirely the same as those determining success as a young officer. One obvious difference might be the factor of motivation. Many of these officers may have been bright and capable, yet only interested in fulfilling their minimal military obligation. However, discussion at this point would be speculative rather than truly data-based.

In conclusion, it appears somewhat remarkable that OEC measures given so early after entry into the Army were able to measure something of what distinguishes a future career officer from a non-careerist. Given a few more years we will be able to determine how well these variables can discriminate among the successful and the "super-successful," i.e., those officers who become colonels and generals rather than retiring as lieutenant colonels. Perhaps the best is yet to come.

References

Helme, W. H., Willemin, L. P. and Grafton, F. C. Dimensions of leadership in a simulated combat situation (Technical Research Report 1172). Arlington, Va.: U. S. Army Behavior and Systems Research Laboratory, July 1971. (NTIS No. AD-730-315)

Helme, W. H., Willemin, L. P. and Grafton, F. C. Prediction of officer behavior in a simulated combat situation (Research Report 1182). Arlington, Va.: U. S. Army Research Institute for the Behavioral and Social Sciences, March 1974. (NTIS No. AD-779-445)

Willemin, L. P. Prediction of officer performance (Technical Research Report 1134). Washington, D. C.: U. S. Army Personnel Research Office, March 1964. (NTIS No. AD-600-038)

McKenzie, Robert C., US Office of Personnel Management, Washington, DC.
(Thurs. P.M.)

The Problem of Range Restriction in Test Validation

This paper examines some of the legal arguments advanced over the years with respect to the interpretation of the validity coefficient. The issue as to whether or not the validity coefficient is of a magnitude large enough to have "practical significance" is given considerable attention. The Equal Employment Opportunity Commission has argued that the validity coefficient is usually too low to have practical significance. Educational testing Service (ETS), however, has contended that the validity coefficient does have practical significance but it should be corrected for restriction in range. The Department of Justice claimed that range restriction correction formulas cannot be used because the assumptions underlying them cannot be met. The Department of Justice in its defense cited the Division 14 Principles for Validation and Use of Personnel Selection Procedures, which asserts the desirability of having validation samples be as similar as possible to the applicant pool. Thorndike's range restriction correction formulas and their underlying assumptions are carefully reviewed in this paper. These formulas are applied to data on actual selection instruments to obtain the estimated true validity coefficient that would be obtained if the valdity coefficient was based on the total applicant population. Also, criterion referenced tests are discussed and suggested as viable alternatives to norm referenced tests, along with factors contributing to criterion biases.

The Problem of Range Restriction in Test Validation

Robert C. McKenzie
U.S. Office of Personnel Management
Washington, D.C. 20415

The opinions expressed in this paper are those of the author and do not reflect official policy of the Office of Personnel Management.

Finding ways to improve selection systems have been at the forefront of personnel psychology for many years. Very sophisticated selection systems have been developed over the years but most have lacked good validity. The most recent descision of the Justice Department with respect to the Professional Administrative Career Exam (P.A.C.E.) has called for more tests to be developed which have higher validity and less adverse impact on minority group members.

Boldt (1977) has pointed out the practical significance that the validity coefficient plays in the legal process through the EEOC guidelines. The EEOC guidelines asserted that test validity must not only be statistically significant but of a magnitude of which to suggest that the benefits obtained from using a particular selection device is worth the trouble of using it. Boldt has also pointed out that the educational testing service has asserted that the correlation coefficient was the appropiate statistics but should be corrected for restriction in range.

The Department of Justice claimed that correlations between selection devices and criteria should not be corrected for range restriction because the assumption of homoscedasticity, linearity, and normality cannot be met. This question, however, raises an important question as to which validity coefficient is the best, the uncorrected or corrected. The purpose of this paper is to shed some light on some of the problems associated with validity coefficients and provide an overview of three main cases of correcting for range restriction in testing.

Generally when a test is administered, it is administered to applicants who walk into a testing center and take a test. However, in some situations, the tester may have control over the applicant pool, but this situation very rarely occurs. Since the total number of people who take the test makeup the applicant pool, the normative information which is obtained from the sample should be based on the total applicant pool. Very seldom, however, are validity coefficient provided for the entire applicant group to whom the test is administered. Consequently, validity coefficients can only be obtained for those people who are selected on the job. When this occurs, the range of test scores becomes restricted, and the sample ceases to become a representative sample of the general population of applicants and thus cannot be generalized to the total applicant population.

Range restriction, as applied to test scores, is a general term which means that the test scores for a particular group are concentrated in only a portion of the possible range of scores (Kaufman, 1972). Groups that are restricted in range have smaller standard deviations than groups that are not restricted. Also, when test scores are restricted in range, correlation between a test and a criterion will be lower than the scores for the unrestricted groups.

When a high standard of selectivity is employed, the effect of range restriction on the resulting validity coefficients becomes even more profound. Thorndike's (1949) frequently cited wartime study is a good illustration of this relationship. In this study, a battery of tests to predict success in pilot training was give :o a large group of men as a part of an Army Air Force Aviation Psychology Program. In using the strict selection standards that were in effect toward the end of World War II, only 13% of these men would have qualified to enter pilot training on the basis of test scores. However, all of the men, regardless of their test scores, were allowed to enter training for experimental purposes. Table 1 below illustrates the correlation coefficients between test scores and the criterion for the total and qualified groups entering training.

TABLE 1

	Correlation with Criterion	
	Total Group	Qualified Group
Predictor	(N = 1036)	(N = 136)
Pilot Stanine (Composite Score)	.64	.18
Mechanical Principles Test	.44	.03
Complex Coordination	.40	-.03
Instrumental Comprehension Test	.45	.27
Finger Dextrity Test	.18	.00
Ceneral Information Test	.46	.20
Arithmetic Reasoning Test	.27	.18

From Thorndike, 1949, page 171

These results show the effect that can occur on validity coefficients when restricted and unrestricted groups are compared. According to table 1, the composite aptitude score has the highest correlation (.64) when the total group is taken into consideration, but a substantially lower correlation (.18) when the group is restricted. Judging by the qualified group alone, the Complex Coordination Test and the Mechanical Principles Test were among the worst predictors. However, for the total group, both the Mechanical Principles and Complex Coordination Tests were among the best predictors.

In order to make practical use of validity statistics for a restricted group, it is necessary to have statistical correction procedures to estimate what validity coefficients would have been obtained if it had been possible to obtain test and criterion data from a representative sample of those to whom the selection device was applied. Thorndike (1949) discussed three types of correction procedures for range restriction, Case I, Case II, and Case III.

Case I occurs when there is some degree of truncation on the criterion (Schmidt and Hunter, 1977; Gullisksen, 1950). In most practical situations, a test or multivariables such as exceptionally good recommendations and exceptionally good academic records are used for selection. However, there may also be situations wherein the criterion itself is used to select employees. For example, a manufacturing company may wish to develop a test to predict job performance and may accept all applicants regardless of their test scores, but weed out the bottom 50% whose performance falls below a given minimum standard. The Case I model is rarely used in validity studies because very seldom is selection made on the basis of the criterion. Selection is almost always made on some type of test. Nevertheless, this case can be put into practice by selecting on the criterion at a point which corresponds to the cutting score where selectees have been proven to be successful. Figure 1 is an illustration of selection occurring on the criterion.

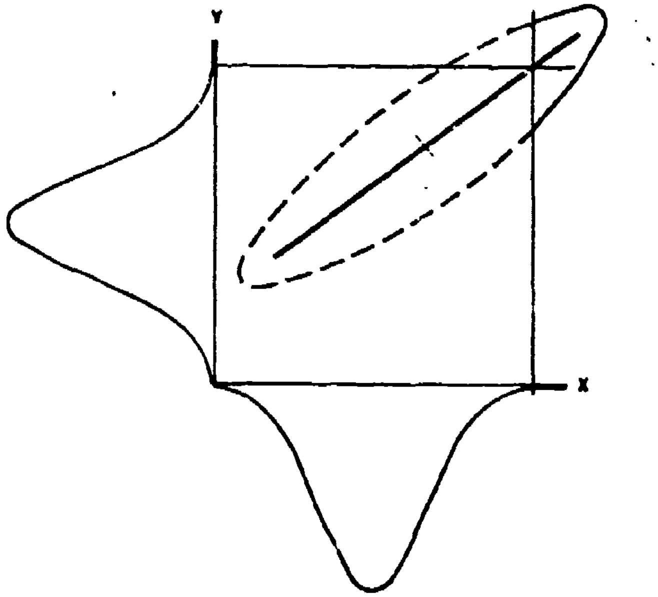


Figure 1. is an illustration of selection occurring on the criterion (Y).

Let Y represent the criterion, and X the predictor. Suppose that the selection ratio is set at .05, then if the assumptions of linearity and homoescedacity hold, setting a cutting score at the top 5% on the criterion will correspond to the top 5% if the test were used to select employees. The dotted and solid line in figure 1 shows the ellipse for the total unrestricted group, and the solid line represents the ellipse for the restricted group. The higher the selection ratio the more severe the restriction in range. Essentially, This is a case of regression of X on Y, instead of the more typical case of regression of Y on X. While this approach tends to be somewhat costly, it does offer an alternative to written tests.

Case II

Case II situation occurs when there is direct truncation on the predictor variable. This model leads to underestimates of the effect of range restriction, if selection is partly on the basis of the criterion as well as the predictor. However, the Case II formula provides slight overestimates of the corrected coefficients when truncation on the test is not perfect. For example, when some applicants below the cutoff score are selected, or selection is made on the basis of the sums of scores on two or more tests rather than a single test (Schmidt and Hunter, 1977).

Table 2 was developed by using the Case II formula (appendix A), and facilitates the determination of R_{12}, the estimated correlation between predictor and criterion in an unrestricted sample.

TABLE 2

Validity Coefficients for Unrestricted Group (R_{12}) Estimated From Values for Restricted Group (r_{12})

$\frac{S_1}{s_1}$	$r_{12} \rightarrow$	.10	.15	.20	.25	.30	.35	.40	.45
1.25		.12	.19	.25	.31	.37	.42	.48	.53
1.50		.15	.22	.29	.36	.43	.49	.55	.60
1.75		.17	.26	.34	.41	.48	.55	.61	.66
2.00		.20	.29	.38	.46	.53	.60	.66	.71
2.50		.24	.35	.45	.54	.62	.68	.74	.78
3.00		.29	.41	.52	.61	.69	.75	.79	.83
4.00		.37	.52	.63	.72	.78	.83	.87	.90
5.00		.45	.60	.71	.79	.84	.88	.91	.93
10.00		.71	.83	90	.93	.95	.97	.97	.98

From Kaufman, 1972, page 6.

$\frac{S_1}{s_1}$ is the ratio of the standard deviation of the unrestricted group to the standard deviation of the restricted group on the predictor test; r_{12} is the actual obtained validity coefficient for the unrestricted group. The unrestricted coefficient is found by looking up the ratio of the unrestricted standard deviation to the restricted coefficient at the top of the table. For example, if the ratio of the standard deviation was found to be 1.25, and the restricted validity coeficient was .25, then table 2 would be entered as follows:

$\frac{S_1}{s_1}$ = 1.25, r_{12} = .25, which results in an unrestricted validity coefficient of (R_{12}) .31.

The magnitude of a pearson product moment correlation coefficient is directly related to the standard deviation of the two variables being correlated. A reduction in either or both of the standard deviations will lower the correlation coefficient between the two variables.

Case III

The third case of range restriction occurs when there is truncation on a third variable. Gulliksen (1950) distinquishes between explicit selection and incidental selection. Explicit selection is defined as direct selection occurring on the basis of a given variable (test), and incidental selection occurs when there is indirect selection occurring on the basis of a given variable (criterion) or another test which is highly correlated with the explicit variable. For instance, suppose a researcher is interested in trying out a new test (Y) to see how well it predicts job performance, and the scores are available for the first test (X) which was used for selection. Selection is then incidental with respect to test Y (new test) because of the high correlation between the new test and the initial test, thus selecting on the first test is basically the same as selecting on the second test. Restriction in range occurs on the new research test because there will be some applicants who passed the first test who will not pass the second test. The Case III situation is more practical than the Case I situation and is very frequent in concurrent validity studies.

Thorndike (1949) discussed another Case III in which one of the correlations used in the above formula maybe available for the total group rather than the restricted group. For example, a research test may have been given to a general unselected group and its correlation with the score on the selection test is based on this group (see appendix A for Case III formula).

These formulas are basically used for pearson-product moment correlations based on continuous variables; however, if the formula for biserial correlation is to be used in restricted groups, the distribution of traits underlying that dichotomy must be normally distributed in the restricted groups (Thorndike, 1949).

Campbell (1976) and Linn (1968) have pointed out that attempting to estimate reliabilities and validities in the appropriate population can pose a serious problem. For instance, they argue that the assumptions of linearity and homoscedasticity (equality of conditional variances) may easily be violated. For example, in a bivariate distribution, much less variability tends to be exhibited at the extremes as opposed to the middle, and so rather than being linear, the standard score of the regression line tends to be steeper at the ends than in the middle. The violation of the homoscedasticity assumption tends to inflate the corrections, while the departure from linearity tends to deflate it. Linn (1968) has shown that by correcting correlation coefficients as if the test (x) is the explicit selection variable, results into the formula overcorrecting when the correlation between x (test 1) and z (criterion) is low and under-correcting when the correlation is in the middle range. The undercorrection and overcorrection phenomena becomes more profound as the degree of actual selection on z (unavailable predictor) becomes extreme. This relationship is shown in figure 2 below which was abstracted from the Linn (1968) report.

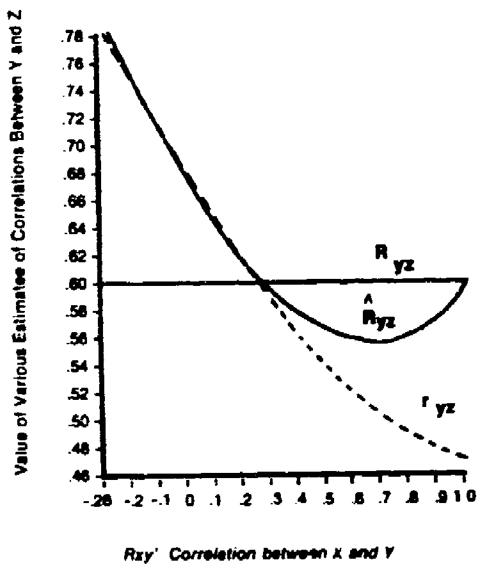


Figure 2. An illustration of the effects of correction of range as a function of the correlation between the explicit and implicit predictors.

Let x be the explicit selection variable (test used for selection), y the implicit selection variable (a second test), and z the criterion but also subject to implicit selection. Then the correlation between y and z in the selected group (r_{yz}) is a function of the correlation between x and y in the unrestricted group (R_{xy}). The broken line in the figure 2 shows the value of r_{yz} in the restricted population as a function of R_{xy} for values of R_{xy}. If y is treated as an explicit selection variable and r_{yz} is corrected for homogeniety of variances of y in the restricted population, the corrected values of $\hat{R}_{yz}$ will still be an underestimate of R_{yz} as long as x and y are reasonably correlated.

A number of studies have shown that the assumptions of linearity, normality and homoscedasticity are very rarely violated. Sevier (1957), using sample sizes ranging from 105 to 250, has shown that out of 24 tests, only one violated the assumption of linearity, and one out of eight violated the assumption of homoescedasticity. Ghiselli and Kahneman (1962) examined 60 aptitude variables in a sample size of 200 and showed that 40 percent of the variables departed significantly from the linear-homoescedastic model. However, ninety percent of these variables held up when cross validated. Tupes (1964) re-analyzed the Ghiselli and Kahneman studies and found that only 10 percent of these relationships departed from linearity at the .05 level.

Schmidt (1978) has pointed out that the assumptions for range restriction correction formulas in real data will be violated only infrequently. However, even if they are violated (e.g., lower conditional criterion variances in the restricted group), there is every reason to believe that the amount of induced bias will be small relative to massive bias induced by failure to correct. This point is illustrated by re-examining the figure abstracted from the Linn (1968) article[1]. Campbell (1976), however, reviewed the Linn article and came to the same conclusion as did Linn. He then further emphasized that these kinds of consideration make using correction

1. As noted by Schmidt (1978), the Linn article is very confusing because figure 2 has been called figure 1, and vice versa. However, by considering the example on which figure 1 is based, a very serious violation of the assumption is made. In this example, restriction has been on a third variable (X) and so the appropiate model is Thorndike's (1949) Case III formula. However variable X is unknown and thus the correct correction formula cannot be used. The incorrect Case II model is used instead, which assumes direct restriction on the predictor.

formulas rather risky business. In almost every real situation an estimate of the population parameter will most likely be biased in some respect, except, in clear and straightforward situations.

Criterion-Referenced Tests

Perhaps a better way to insure that selection systems have validity is to make sure that the test is an adequate sample of the job content. Criterion-referenced tests although less sophisticated than norm-referenced tests provide better samples of the job domain. These tests are designed to measure performance relative to a criterion standard. That is, an employee's test performance is compared to an established criterion performance standard. Normative data as to how well one employee is performing relative to another is not necessary. The job is usually simulated with 90% accuracy, so that the performance on the test is basically the same as the actual performance on the criterion. Criterion-referenced test are basically the same as performance tests.

Job performance tests such as work sample tests are especially applicable as use for criterion-referenced measures (Buck, 1975). These tests like criterion-referenced tests, must sample the job domain accurately and also must measure the employee's ability to perform critical job tasks. As Buck (1975) has pointed out, criterion referenced tests are often confused with criterion related tests. A criterion related test implies that there is some kind of statistical relationship between a test and some measure of job performance (i.e., supervisory ratings, self ratings, peer ratings, etc.). While on the other hand, criterion-referenced tests refer to the minimal acceptable level that an applicant must meet in order to achieve a mastery level on the job. Essentially, criterion-referenced tests are used to determine who is qualified, they do not measure variability in performance. Popham and Husek (1971) contend that criterion-referenced tests are only suitable when there is no constraints on how many people are to be selected for a particular job. However, when there are constraints, norm-referenced tests are usually more appropiate for selection. Like norm-referenced tests, criterion-referenced tests can also be used when the applicant pool is exceptionally large by setting the cutting score above the minimall qualifying standard. For example, if the minimally qualifying standard cut score is set at 70%, it can be set at 90% which would insure the adaptation of a treatment that would select superior employees. The advantages of this aproach of testing over norm-referenced tests is the accuracy with which the job domain is sampled and the fact that less complex tests can be developed which in turn would give better measures of validity.

Reference Notes

1. Schmidt, F.L. Personal communication, January 1978

References

Boldt, R.F. Robustness of range restriction in court, Unpublished report, 1978.

Buck, L.S. Use of criterion-referenced tests in personnel selection: A Summary Status Report. Technical Memorandum, 1975 6

Campbell, J.P. Psychometric theory. Handbook of Industrial and Organizational Psychology, 1976.

Gullisksen, H. Theory of Mental Tests. New York: McGraw Hill, 1956.

Popham, W.J., & Husek, T.A. Implications of criterion-referenced measurement. In W.J. Popham (ED.), Criterion-referenced measurement (An introduction). New Jersey: Educational Publication, Inc., 1971.

Kaufman, A. S. Restriction of Range: Questions and Answers. Test Services Bulletin, The Psychological Corporation, 1972.

Levin, J. The occurence of an increase in correlation by restriction of range. Psychometrika, 1972 37, 93-97.

Linn, Robert L. Range restriction Problems in the use of selected groups for test validation. Psychological Bulletin, 1968, 69, 69-73.

Schmidt, F.L. & Hunter, John E. The development of a general solution of the problem of validity generalization. Journal of Applied Psychology, 36, 1977.

Thorndike, Robert L. Personnel Selection. Test & Measurement Techniques, New York, 1949.

Ghiselli, E.E. & Kahneman, D. Validity and non-linear heteroscedastic models. Personnel Psychology, 15 , 1962,

Sevier, F.C. Testing the assumptions of underlying multiple regression. Journal of Experimental Education, 1957, 25, 323-330.

Tupes, E.C. A note on "Validity and Non-linear heteroscedastic models". Personnel Psychology, 17, 1964.

Thorndike's Case I formula:

$$R_{xy} = \sqrt{1 - \frac{S_y^2}{s_y^2}\left(1 - r_{xy}^2\right)}$$

The Case I formula is primarily concerned with the correlation between x and y. Where R_{xy} is the correlation between a test (x) and a criterion (y) in the unrestricted group, Sy is the standard deviation of the criterion in the unrestricted group, sy is the standard deviation of the criterion in the restricted group.

Thorndike's Case II formula:

$$R_{xy} = \frac{r_{xy}\frac{S_x}{s_x}}{\sqrt{1 - r_{xy}^2 + r_{xy}^2\frac{S_x^2}{s_x^2}}}$$

A special case of Case III when the validity coefficient is known for the unrestricted group:

$$R_{xy} = \frac{r_{xy}\sqrt{1 + R_{xz}^{2}\left(\frac{s_z^{2}}{S_z^{2}} - 1\right)} + R_{xz}\, r_{yz}\left(\frac{S_z}{s_z} - \frac{s_z}{S_z}\right)}{\sqrt{1 + r_{yz}^{2}\left(\frac{S_z^{2}}{s_z^{2}} - 1\right)}}$$

The terms are define in the Case I formula.

Thorndike's Case III formula:

$$R_{xy} = \frac{r_{xy} + r_{xz} + r_{yz}\left(\frac{S_z^2}{s_z^2} - 1\right)}{\sqrt{\left[1 + r_{yz}^2\left(\frac{S_z^2}{s_z^2} - 1\right)\right]\left[1 + r_{xz}^2\left(\frac{S_z^2}{s_z^2} - 1\right)\right]}}$$

These symbols are analogous to the Case II formula, with the exception of the third variable being correlated. R_{xy} is the correlation between the explicit variable (test) and the implicit variable (a new research test) for the unrestricted group, r_{xy} is the correlation between the explicit variable and the implicit variable for the restricted group, r_{xz} is the correlation between test X and the criterion Z for the restricted group, r_{yz} is the correlation between implicit variable (new research test) and the implicit variable (criterion Z for the restricted group, and S_z^2 and s_z^2 are the unrestricted and restricted group variances for the criterion Z.

AD P001353

Determinants of Readiness for Reenlistment in the Federal Armed Forces

Adelheid Meissner and Klaus J. Puzicha
Federal Armed Forces Office
(Germany)

Summary

In december 1980 a representative sample of soldiers of all three forces of the German Federal Armed Forces (ranks: private to sergeant) was questionned. Topic of this research work was the analysis of determinants of job satisfaction and readiness for reenlistment. Controlling for the variables "force" and "status" (compulsory condition vs. enlistment for two years vs. enlistment for three or four years vs. enlistment for five years or longer) we have analysed the most important determinants by multiple regression.

The main determinants of job satisfaction and readiness for reenlistment are: affinity for military concerns, perception of the military superiors and the conditions of work. Readiness for reenlistment in particular has additional social determinants: the attitude of the wife respectively the girlfriend and personal contacts with the union-leader.

1 Problem

The German Federal Armed Forces contain of about 500.000 soldiers. Corresponding to a NATO-agreement it is demanded that 55 per cent of them should be enlisted men. But in fact at present there are more than 50 per cent draftees who serve with the Armed Forces. This overhang compensates a lack of corporals: So at last the soldier under compulsory conditions has to undertake the function of a corporal. (1)

This was the situation till the year 1978. Meanwhile the personnel development has changed:

- In the Army and in the Navy there is a shortage of middle- and long-term volunteers, i.e. of soldiers who enlisted for three and more years. In these two forces the lack can be compensated by increasing quota of short-term volunteers (enlistment for two years). But this compensation is only a quantitative one; there still remains the problem that even young soldiers with short enlistment are overtaxed.
- The personnel situation in the Air Force is characterized by a considerable deficiency of short-term enlisted men (2 years) and long-term volunteers (8 years and longer) on the one hand and an overhang of soldiers with medium enlistment (3 to 7 years) on the other hand.

One problem of the German Armed Forces in the eighties is how to fit the needs of enlisted men in a satisfactory way. Principally there might be three possibilities:

- enlistment of volunteers
- change of status from draftee to volunteer
- reenlistment of voluntary soldiers.

During the last years the requirements of volunteers are covered by the three mentiond facilities to a thired each. (2)

The present research concerns with the problem of deficient readiness for reenlistment respectively enlistment by draftees. Both aspects we have summarized as READINESS FOR REENLISTMENT. We consider this problem as an indicator of JOB SATISFACTION which was the other main subject of our investigation.

The following survey shows the design we have applied to analyse these two aspects.

DEPENDENT VARIABLES
(Criteria)

relevant for the individuum	relevant for the institution
satisfaction with...	
- ...superiors	- readiness for reenlistment
- ...conditions of work	- efficiency as multiplicators
- ...comrades	- extrinsic motivatibility
- ...payment	

INDEPENDENT VARIABLES
(Predictors)

more subjective ones	more objective ones
- personality traits	- biographic data
- perception of the superiors	- objective aspects of the military job
- perception of the conditions of work	
- perception of the comrades	

CONTROLLING VARIABLES

- forces (Army vs. Airforce vs. Navy)
- status (service under compulsory conditions vs. enlistment for two years vs. enlistment for three or four years vs. enlistment for five years or longer)

2 Methods

At the end of 1980 an investigation concerning the mentioned problems was carried out by the German Armed Forces' Psychological Research Institute Bonn. On the basis of randomized sample selection 1500 soldiers were interrogated postally by a pretested questionnaire. The sample was representative for the ranks from private to sergeant - containing of both draftees and volunteers.

Description of sample

	Army	Airforce	Navy
servicemen	643	438	152
corporals	55	-*	34
sergeants	121	-*	116

The quota of recoil was about 42 per cent.

*Because of a technical error in the Airforce there were only servicemen available.

To analyse the determinants of job satisfaction respectively readiness for reenlistment multivariate methods such as multiple regression analysis seemed adequate. For criteria (that means our dependent variables) we have taken the following dimension which we have got by factor analysis of some selected measuring instruments:

- Satisfaction with the superiors
- Satisfaction with the conditions of work
- Satisfaction with the comrades
- Satisfaction with the payment
- Readiness for reenlistment
- Efficiency as multiplicators
- Extrinsic motivatibility

Some of these dependent variables have been used for predictors, too. In the following all applied predictors are specified.

1	Rank	
2	Seniority	
3	Futural time of service	objective aspects of the military job
4	Living in barracks	
5	Time of reversion	
6	Shifting	
7	Level of education	
8	Failing remove	
9	Confession	
10	Family status (married)	
11	Distance from wife/girlfriend	
12	Attitude of wife/girlfriend towards reenlistment in the Armed Forces	objective aspects of biography
13	Unemployment of colleagues or friends	
14	Satisfaction with the educational and professional level	
15	Satisfaction with the superiors	
16	Authoritarian superiors	
17	Number of contacts with the union leader	perception of the superiors
18	Foreign-determination	
19	Satisfaction with the conditions of work	
20	Adequate function	perception of the conditions of work
21	Burden by special missions and long-term working hours	
22	Idleness	
23	Satisfaction with the comrades	perception of the comrades
24	Alcohol and missing comradeship	
25	Alcohol consumption	
26	Satisfaction with the payment	–
27	Expectations before joining the Armed Forces	
28	Prejudices towards the Armed Forces	
29	Affinity for military concerns	personality traits
30	Unpolitical attitude	
31	Readiness for leadership/dominance	
32	Need for achievement	
33	Immobile passivity	

3 Results

In a first descriptive view of results it seems opportune to give a kind of inventory, i.e. such questions as "Which groups of soldiers are ready to reenlist, which are not?" or "How can one improve the readiness for reenlistment?"

Readiness for reenlistment in several groups					
Groups		"I want to become professional"	"I will reenlist"	"I am still indecisive"	"I will not reenlist"
Army	%	4	5	14	77
Air Force	%	2	4	11	83
Navy	%	12	4	18	66
servicemen	%	2	4	12	83
corporals	%	5	11	25	59
sergeants	%	17	7	18	58
draftees	%	1	3	7	90
enlisted men for 2 years	%	1	5	20	75
enlisted men for 3 or 4 years	%	6	9	29	56
enlisted men for 5 years or longer	%	24	7	19	50

Four of five Air Force soldiers, three of four Army soldiers and two of three Navy soldiers don't want to extend their time of service voluntarily.
Soldiers with a higher status show more readiness for reenlistment or for becoming professional.

Our results made clear that there are considerable differences between soldiers of the three forces with regard to job satisfaction and readiness for reenlistment. Likewise the results for the various status groups were not comparable. Therefore we have made all regression analyses separate for Army, Air Force and Navy as well as for the groups:

- service under compulsory conditions (draftees)
- enlistment for two years
- enlistment for three or four years
- enlistment for five years or longer

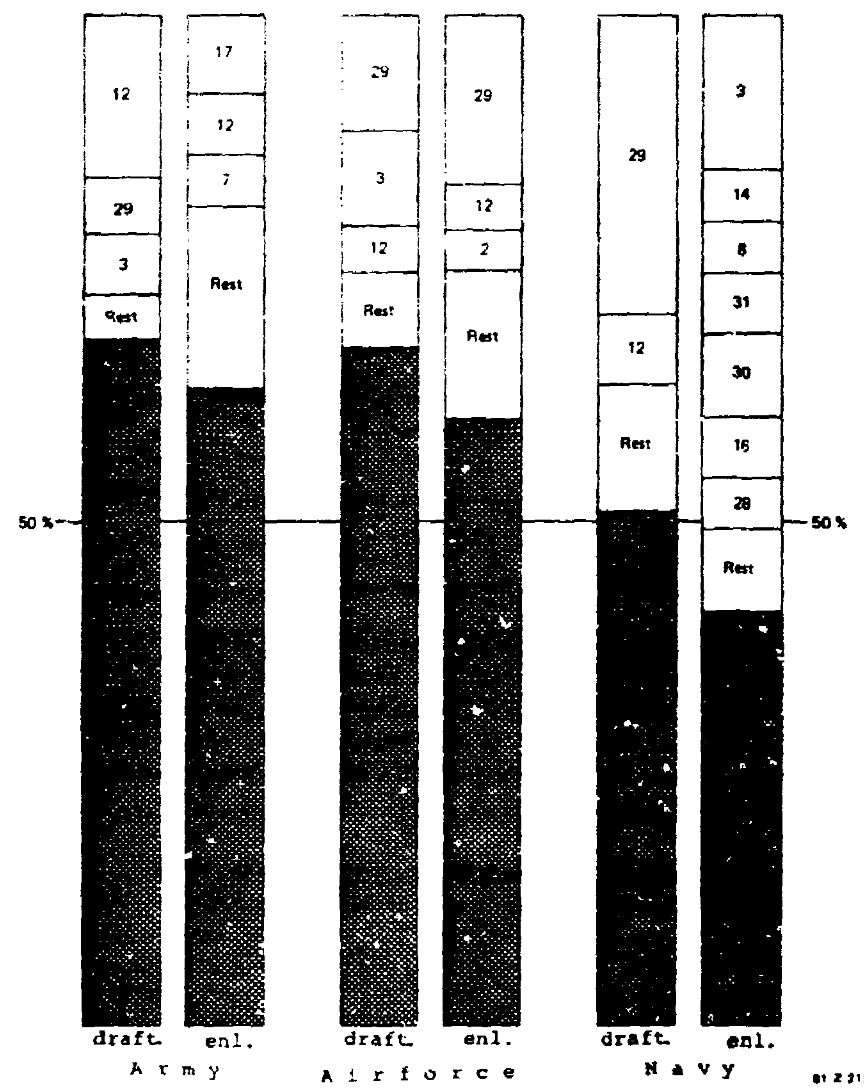
Though we cannot present all these single results here we will try to select some characteristic tendencies. The determinants of readiness for reenlistment are shown at the following diagrams. The hatched parts of the columns represent the portion of not explainable variance. The numbers stand for the various predictors. Some important ones are specified here:

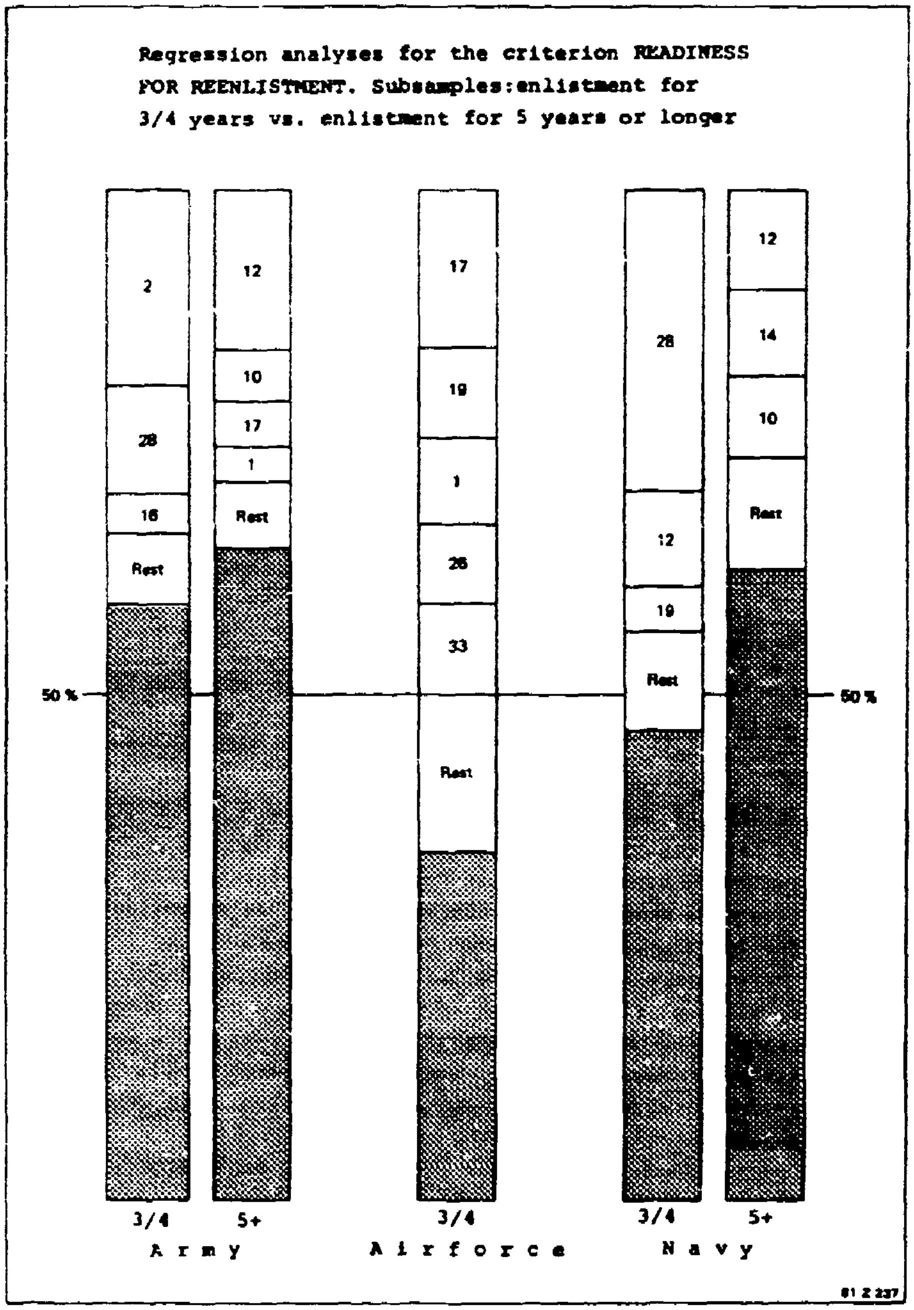
29 affinity for military concerns
12 attitude of the female partner towards reenlistment
17 number of contacts with the union leader
3 futural time of service (the more distant the term of a decision - reenlistment or not - the greater the inclination to do it)
2 seniority
28 prejudices towards the Armed Forces

We can see that for draftees and for short-term volunteers affinity for military concerns (29) and attitude of the girlfriends (12) are most relevant. For the subsamples of long-term volunteers attitude of the female partner (12), contacts with the union leader (17), prejudices towards the Armed Forces (28) occupy the first ranks.

In a further step we have summarized the ten most important predictors of the three criteria concerning reenlistment.

Regression analyses for the criterion READINESS FOR REENLISTMENT. Subsamples: draftees vs. enlistment for 2 years




Regression analyses for the criterion READINESS FOR REENLISTMENT. Subsamples:enlistment for 3/4 years vs. enlistment for 5 years or longer
2
28
16
Rest
12
10
17
1
Rest
17
19
1
26
33
Rest
28
12
19
Rest
12
14
10
Rest
50 %
50 %
3/4
5+
Army
3/4
Airforce
3/4
5+
Navy
81 2 237

Average relevance of the predictors for the three criteria READINESS FOR REENLISTMENT, EFFICIENCY AS MULTIPLICATORS and EXTRINSIC MOTIVATIBILITY
1 Affinity for military concerns
2 Prejudices towards the Armed Forces
3 Attitude of wife/girlfriend towards reenlistment
4 Number of contacts with the union leader
5 Satisfaction with the conditions of the work
6 Satisfaction with the superiors
7 Adequate function
8 Distance from wife/girlfriend
9 Satisfaction with the payment
10 Seniority

Corresponding to the high portion of conscripts in our sample the emotional dimensions "affinity for military concerns" and "prejudices towards the Armed Forces" have the most important relevance. Next ranks are occupied with social influences. They show the importance of reference groups for reenlistment.

The attitude of the female partner and the frequency of contacts with the union leader determine the decision in favour of or against a reenlistment.

In comparison with an american study (3) we have analysed how far the readiness for reenlistment is dependent on some certain material stimuli. We have called this "extrinsic motivatibility". To measure this nine more or less realistic stimuli as potential motivation for reenlistment were given to the soldiers. If only asking for their general attitude 72 per cent of them spontaneously say that they will not reenlist. But confronting them with these "extrinsic" incentives there are only still 14 per cent left, who would reenlist by no means. In other words: 58 per cent of all soldiers were ready to revise their original intention; they were more or less venal.
Yet not all nine stimuli are likewise effective for improving the readiness for reenlistment.

What kind of conditions promise to be most effective?	
Without regard to realization:	
more favorable leisure and compensation of overtime	rank 1
option of garrison and co-determination of removal	rank 3*
certain preferment	rank 3*
better payment	rank 3*
guarantee of garrison near residence	rank 5
obligatory promises of development	rank 6
With consideration to the most realistic measures	
obligatory promises of development	rank 1
certain preferment	rank 2
more favorable leisure and compensation of overtime	rank 3.5*
payment of an attractive enlistment-premium	rank 3.5*
*These conditions occasionally have the same rank	

By suitable combination of the practicable measures 80 per cent of all soldiers could be addressed.

4 Summary and Discussion

As we assume a close connection between job satisfaction and readiness for reenlistment we have summarized all seven criteria and investigated what kind of predictors are most significant.

Average relevance of the predictors for all seven criteria. Total sample
1 Affinity for military concerns
2 Satisfaction with the superiors
3 Satisfaction with the conditions of the work
4 Prejudices towards the Armed Forces
5 Satisfaction with the comrades
6 Attitude of wife/girlfriend towards reenlistment in the Armed Forces
7 Burden by special missions and long-term working hours
8 Foreign determination
9 Alcohol and missing comradeship
10 Number of contacts with the union leader

We can see that altogether "affinity for military concerns" represents the most important determinant, that means the attitude towards the typical aspects of military jobs as living in barracks, wearing uniform, and living with the principle of command and obedience. The significance of this emotional aspect concerning the nearness or distance to the Federal Armed Forces decreases with the duration of enlistment.

The next important factor of influence is the satisfaction with the superiors. That makes clear once more the necessity of good leadership behavior. That is what the German Federal Armed Forces call "leadership and civic education".
Returning to our first mentioned problem - i.e. a low rate of reenlistments of volunteers respectively of enlistments by draftees - we may ask whether the results of our investigation give any hints for improving that situation.
We have found out that an increase of readiness for reenlistment may be realized by certain "extrinsic" stimuli. Even soldiers who actually are against any reenlistment would relent if offered such permissions. Supposing these stimuli were practicable - would such a "venality" be desirable if there are no positive emotional attitudes of soldiers towards the Armed Forces? From our point of view it is quite possible that attitudes may change and develop so that they are in accordance with the decision that has been made before. That means: after one has decided to reenlist he may look at the military aspects in a more positive way than he had done before; his affinity for military concerns increases.

5 References

(1) Bundesminister der Verteidigung: Weißbuch, 1979. Zur Sicherheit der Bundesrepublik Deutschland und zur Entwicklung der Bundeswehr.

(2) Bundesminister der Verteidigung - Abteilung Personal: Jahresbericht 1979/80

(3) Stephan J. MOTOWIDLO et al.: Reenlistment factors for First Term Enlisted Personnel. In: Proceedings 22nd Annual Conference of the Military Testing Association, Toronto, 1980. p.679-690.

TASK ANALYSIS FOR WEAPONS SYSTEMS TESTERS: SHORTCUT TO PAYDIRT IN INFLATIONARY TIMES

AD P001354

John L. Miles, Jr., J.D.
United States Army Research Institute
for the Behavioral and Social Sciences

The principal effects of inflation on testing of new weapon systems are to compress schedules and thereby reduce the range and scope of system performance which can be assessed. Test planners today must analyze system performance goals and system design and then predict those few areas where testing is most likely to pay off. Testing conducted only in accordance with these predictions can fail to detect significant problems which will show up when the system is eventually fielded. A new approach to conducting task analysis may provide better indication to test designers of where to anticipate problems and, hence, where to apply scarce testing resources. This approach was developed by a tri-service committee of human factors practitioners, and its concepts won an 80% indorsement of other practitioners in government and industry who responded to a questionnaire.

While the battles of consumers against inflation may gain the most media attention, the pernicious effects of inflation are felt as well by weapon system testers. Most frequently these effects are felt on test schedules, in which such budget contraints as "number of rounds of ammunition available" dictate shorter and smaller tests. Test planners are thus faced with the dilemma of whether to reduce the sample size on which important generalizations are to be based or to eliminate whole subtests. When this dilemma is resolved in favor of the statisticians, one of the first subtests to be cancelled is often that of human factors. Historical grounds for this decision appear to include the perceptions that: (1) no Army system has ever been cancelled solely for human factors reasons, (2) the legendary ingenuity of the American soldier will enable him to overcome or at least compensate for disadvantageous design of equipment, and (3) if the electrical and mechanical subsystems can be made to operate with satisfactory reliability, the whole system is probably good enough to go to the field.

There is growing evidence today that those grounds are being undermined. Human factors deficiencies were prominent among the reasons for Congressional cancellation of the Family of Military Engineering Construction Equipment (FAMECE) project, and the prestigous Kerwin and Blanchard report Man-Machine Interface - A Growing Crisis began, "The US Army has a major man-machine interface problem... The problem is severe and will continue to get worse" (p. 1). Thus, at a time when inflation is exposing testing in general and human factors testing in particular to severe cuts or elimination altogether, the criticality of the man-machine interface is increasing. Prudent test planners are today therefore casting about for more efficient means of determining which areas of system testing are likely to have the highest payoff in terms of testing resources invested.

Human factors practitioners (particularly within the Army) have provided ample guidance:

(1) A general life cycle system management model showing the integration of human factors and identifying test and evaluation activities was published in 1980 (Burt, et. al., 1980).

(2) A detailed guide for gathering and analyzing human performance data during developmental testing was issued by the USA Human Engineering Laboratory in 1976 (Berson and Crooks, 1976).

(3) A workbook and a handbook for assessing human performance during operational testing (known as HRTES) were developed by the US Army Research Institute (Kaplan, et. al. 1980).

Although each of those documents has been well received in the professional community, there remains one persistent problem whose solution has eluded the human factors and training community for years and is a necessary prerequisite to efficient determination of high payoff areas for investment of testing resources. That problem is determining and then documenting what the humans in the system have to do (and, therefore, what training they have to receive) to make it work properly. The means most often used for this undertaking is task analysis.

Task analysis is of course not new, and its origins may lie with the 1898 work of Frederick Taylor at the Midvale Steel Company. A paper by Hays (in press) contains a recent review of the status of task analysis, and another by Berry (1979) contains a succinct summary of the problem:

> Modeling of a human-machine system for whatever purpose, requires that all significant events occurring in the system be described. Task analysis, standard in the repertoire of the human factors engineer, is the technique generally used to describe the activities performed by the human components, or operators. Unfortunately, there is no agreement on the vocabulary or structure to be used in making these descriptions.
>
> Within the past 25 years, at least a dozen task classification schemes, or taxonomies, have been proposed. Even definition of the term task is not universally agreed upon, and this definition is critical because it strongly influences the terms, units and general flavor of the final taxonomy (p. 1).

The difficulties with task analysis are well-known to developers of military systems. "Happy hour" conversations often contain stories of defective task analyses, but virtue alone is not enough: during the late 1960s and early 1970s there was a joint German-American development program for a main battle tank. The tank analysis for that system when delivered was 32-inches thick and no one was able to use it (Brogan et. al, 1981, p. 262). Attempts were begun in the late 1970s in the Department of Defense to solve the problem of task analysis. As General Becton explained the problem to the Army's Vice Chief of Staff,

> A proper basis for both human engineering and human factors to pursue their goals is a task analysis. This is an analysis of those specific tasks an operator must perform to operate the system. A valid task analysis provides a logical basis for human engineering design, training program development, training device requirements, and other management considerations such as SQTs and MOS prerequisites. Too often in the development of Army equipment systems, task analyses are not done or are done incompletely. This is a primarily because no military standard currently exists for task analysis, the users of task analysis information have different requirements, multiple formats are used, and the task analysis must be called out as a deliverable data item to be performed (pp. 1-2).

In General Becton's view, what was needed was a military solution to a technological problem.

An intrepid group of military and civilian trainers, testers and human factors specialists drawn from all three armed services began work on the problem in late 1978. By June of 1979 the group (officially designated as the Test and Evaluation Subgroup of the Department of Defense Human Factors Engineering Technical Advisory Group, but more commonly known as the "T&E SubTAG") had reviewed the task analysis programs in all three services, developed its own task taxonomy (see Figure 1),

<u>Mission:</u> What the man-machine system is supposed to accomplish.
<u>Scenario/conditions:</u> Categories of particular factors or constraints under which the system will be expected to operate and be maintained.
<u>Function:</u> A broad category of activity performed by a man-machine system.
<u>Job:</u> The combination of all human performance required for operation and maintenance of <u>one</u> personnel position in a system.
<u>Duty:</u> A set of operationally-related tasks within a given job.
<u>Task:</u> A composite of related activities (perceptions, decisions, and responses) performed for an immediate purpose, written in operator/maintainer language.
<u>Subtask:</u> Activities (perceptions, decisions and responses) which fulfill a portion of the immediate purpose within a task.
<u>Task Element:</u> The smallest logically and reasonably definable unit of behavior required in completing a task or sub-task.

Figure 1. Task Taxonomy from Proposed Military Standard

and outlined a proposed military standard on task analysis (Miles, 1979). A questionnaire was subsequently developed at Edwards AFB, California and sent by name to over a hundred practitioners of human factors and training in government and industry to obtain their reactions to both the proposed taxonomy and the idea of having a military standard serve as arbitrament in this area. To the utter astonishment of the subTAG, more than 80% of the survey respondents agreed with the proposal (which may be an indication that people other than Army generals are tired of chaos in this area). Some minor revisions were then made to the conceptual scheme, and the draft military standard was prepared (Zavala, 1980).

The drafters of the standard had two innovative goals. First, they wanted to use the same task data base for all of the specialty programs (design, training, test and evaluation, manning and workload) which traditionally have required task analysis. This was both for purposes of economy (it's hard enough to get a project manager to buy one task analysis -- let alone five) as well as the promotion of coordination among the various specialists concerned with aspects of personnel and training. To accomplish that goal, they established two resevoirs of data -- "task inventory" which is rigidly controlled by the task taxonomy, and "supporting data" which is everything else in whatever format that may be needed to insure the accuracy or validity of the task analysis. Second, they tried to reach a new height of specificity and flexibility -- the former so that a contractor would know with precision in every case what the desired task analysis should look like, and the latter to give the government maximum freedom in terms of level of detail while still preserving all of the controls inherent in the proposed standard. This was accomplished in two primary ways: (1) by requiring that every task analysis include two specific levels in the task inventory ("job" and "task") and (2) by prescribing output format but not process. It was reasoned that, with these requirements, both gross and detailed task analyses could be obtained from the same data base for the same system (the latter by adding more of the optional levels of the taxonomy) and that two entirely different systems (e.g., a rifle and a jet aircraft) could still use the same conceptual model of task analysis. A schematic of this model is shown in Figure 2. In this model (and in the draft standard) both the input (at least the task inventory portion) and the output (shown as "Data Item Descriptions" on DD Forms 1664) are carefully prescribed; the process called "task analysis" is not. Therefore the government planner may direct that the contractor use anything from a stubby-pencil manual method to one of the sophisticated ADP-assisted methods such as MOAT (see Helm and Donnell, 1979) -- or even some new technique invented after the standard was written.

Use of this standard should be particularly helpful to testers of military systems. Among the recommendations of Generals Kerwin and Blanchard was "Manpower and skill level specifications and human performance requirements must be developed, stated and used to the same degree materiel specifications have been in the past" (Kerwin and Blanchard, 1980, p. 7). Means for implementing this recommendation were proposed by Kaplan and Crooks (1980) based on integrating the draft task analysis standard with their earlier efforts on HRTES. In those projects where that recommendation is in fact implemented, objective, verifiable human performance criteria will exist in requirements documents which can easily be translated into test design plans and then into detailed test plans.

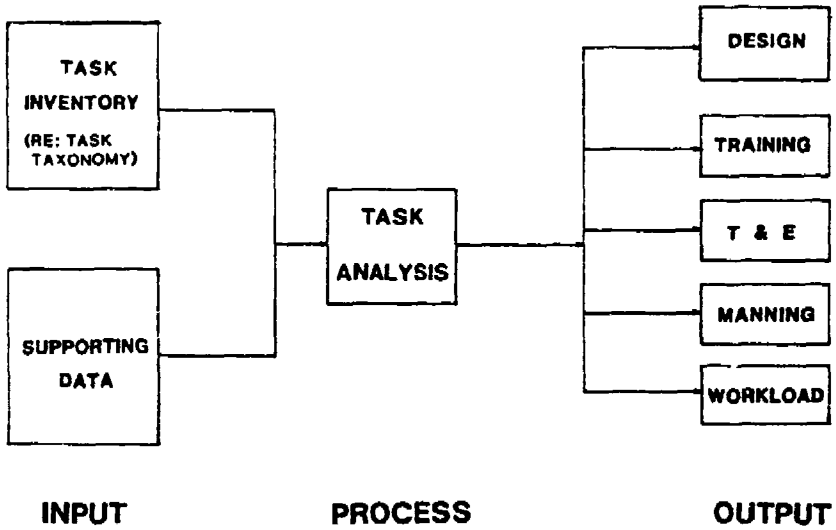


FIGURE 2. SCHEMATIC OF TASK ANALYSIS.

Summary

The draft military standard on task analysis was created to bring both order and standardization to the process of describing and documenting what the humans in a military system are required to do to make it function properly. Its use permits testers of military systems to identify quickly those human performance criteria considered of primary importance in obtaining the forecast level of system effectiveness.

References

1. Becton, General Stanley, Memorandum for Vice-Chief of Staff, United States Army, Subject: Human Engineering Testing in Materiel Development. Falls Church, VA: USA Operational Text and Evaluation Agency, 28 February 1978.

2. Berry, Gayle L., Task Analysis: Prerequisite to Human - Machine Systems Performance Prediction. Ph.D. dissertation. 1979.

3. Berson, B. L., and Crooks, W. H. "Guide for obtaining and analyzing human performance data in a materiel development project." Technical Memorandum 20-76. Aberdeen Proving Ground, MD: U.S. Army Human Engineering Laboratory, September 1976.

4. Brogan, R., Hedge, J., McElroy, K. and Katznelson, Judah. Army-Navy Computer-Aided Human Factors Engineering Course, Lesson No. 23, Virginia Beach, VA: Advanced Technology, Inc., September 1981.

5. Burt, D. L., et. al. Human Factors Engineering in Research, Development and Acquisition. Bethesda, MD: Andrulis Research Corporation, 1980.

6. Hays, Robert T., Training Simulator Fidelity Guidance: The Iterative Data Base Approach. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. In press.

7. Helm. LT W. R. and Donnell, M. L., Mission Operability Assessment Technique: A System Evaluation Methodology. Point Mugu, CA: USN Pacific Missile Test Center Technical Publication TP-79-31, 10 October 1979.

8. Kaplan, J. D., Crooks, W. H., Sanders, M. S., and Dechter, R. "HRTES: Human Resources Test and Evaluation System, Volumes I and II." Working Draft, Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, February 1980.

9. Kaplan, Jonathan and Crooks, W. H. "A Concept for Developing Human Performance Specifications." Technical Memorandum 8-80. Aberdeen Proving Ground, MD: U.S. Army Human Engineering Laboratory, April 1980.

10. Kerwin, Gen. Walter T. and Blanchard, Gen. George S., Man-Machine Interface - A Growing Crisis. Aberdeen Proving Ground, MD: US Army Materiel Systems Analysis Activity Discussion Paper Number 2. August 1980.

11. Miles, John L. Jr., Minutes of the Meeting, Ad Hoc Committee on Task Analysis. Aberdeen Proving Ground, MD: USA Human Engineering Laboratory, 21 June 1979.

12. Zavala, Albert, Draft Military Standard on Task Analysis. Bethesda, MD: Andrulis Research Corp. February 1980.

AN OCCUPATIONAL ANALYSIS OF AIR FORCE BEHAVIORAL SCIENTISTS: PRELIMINARY RESULTS

Lt Col Jimmy L. Mitchell
Mr James B. Keeth, and
Capt Linda A. Wiekhorst

Occupational Analysis Program
USAF Occupational Measurement Center
Randolph Air Force Base, Texas 78150

USAF Behavioral Scientists (AFS 267X) are a relatively small operational group who perform very diverse jobs, primarily in Air Training Command, Air Force Systems Command, and the USAF Academy. Currently, there are about 140 officer positions authorized in this specialty which range in grade from Second Lieutenant to Lieutenant Colonel. An occupational survey of Air Force behavioral scientists and related jobs was conducted during the summer of 1981 to assess the types of jobs being performed, organization of the specialty, career patterns, job interest and satisfaction, and educational requirements for the various types of jobs. The occupational data developed in this study will be used by Air Force Manpower and Personnel Center personnel classification officials to determine if subspecialty designators (shredouts) are needed and to refine the current specialty description (AFR 36-1). The data will also be available for use in the behavioral scientist career development program. This paper will report the initial analysis of the occupational data focusing primarily on the types of jobs performed by members of the specialty. A more complete report of the behavioral scientist occupational analysis project should be available in early 1982.

INTRODUCTION

The United States Air Force (USAF) has a long history of successful behavioral science research. Many of the diverse lines of research grew out of World War II efforts, such as the pilot selection program in San Antonio, and the emerging field of human factors engineering in weapons system design pioneered at the laboratories at Wright-Patterson AFB. Through the years, some of the best known names in psychology have been part of the USAF behavioral science research program, either as civilian employees or while they were military officers. A substantial number of these behavioral scientists have gone on to make major contributions in the academic and applied psychology areas. Most have acknowledged the value of their Air Force experience, but some (cf. Jacoby 1970) have been critical of how the Air Force made use of their talents and abilities. Many of the lines of research and the operational programs which have been developed in the USAF research program require an on-going supply of highly qualified civilian and officer behavioral scientists. The present study represents an attempt to help define the jobs of military behavioral scientists and to examine the diversity of positions (Driskill and Mitchell 1980) within the utilization field. Since one of the most successful operational programs developed by the USAF has been the task-based occupational analysis system (Morsh 1964; Christal 1974), it is particularly fitting to use the Air Force CODAP job analysis technology to study Air Force behavioral scientist officer jobs.

BACKGROUND

The present study emerged out of a number of concerns as to how the USAF behavioral scientist utilization field (AFS 267X) is organized and how the various subspecialties are identified and controlled. In the initial Air Force officer classification scheme of 1954, the field involved four specialty codes—one for Research and Development Officer, and three for Human Resources Research (see Figure 1). Between 1964 and 1976, the AFS 267X area was structured into one specialty, with four subspecialty designators: A = Human Factors Psychologist; B = Experimental Psychologist; C = Personnel Measurement Psychologist; and Z = Other social scientists (including sociologists, physical anthropologists, cultural anthropologists, etc.). In 1976 these shredouts were dropped, leaving one basic AFSC. The staff level specialty (AFS 2616) was continued as a separate code, and grouped with other types of Air Force scientists, such as physicists, chemists, mathematicians, etc.

The authorized grade spread of the 267X field is 2Lt through Lt. Colonel. Those selected for Colonel must give up the 267X designation and become either a Staff Scientist (AFS 2616) or take another O-6 level code (such as Organization Commander 0026, or Plans & Programs Officer 0076). Colonels with a behavioral science background can be assigned to a variety of O-6 positions, and any O-6 can be assigned to the key leadership positions in behavioral science organizations, regardless of their technical backgrounds. This policy is quite consistent with the Air Force "Whole Man" concept of officership in the Line officer corps (versus technical qualification and rank progression in specialty corps such as the medical, dental, legal, chaplain, and other areas).

In 1977, the Air Force personnel system developed Special Experience Identifier (SEI) codes which were used to identify and track high-value individuals, who might be needed for special assignments. Since the subspecialty designators ("shredouts") for behavioral scientists had been deleted in 1976 to provide more assignment flexibility, it was thought that this new SEI system could serve the same purpose, without the assignment limitations inherent in in AFSC shredouts.

During 1977 and 1978, a number of complaints were made about the lack of shredouts, as individuals who did not have the special backgrounds needed to perform specific jobs were assigned to various organizations. The Human Factors positions were particularly vulnerable. Thus, some Human Factors graduates (of programs such as Purdue's special HFE masters under McCormick) were assigned to the WAPS test development program, while one graduate student of Dunnette at Minnesota with a major in motivation theory ended up at the Human Factors Branch at the Flight Test Center at Edwards AFB CA as a human factors engineer.

Such assignment problems and a growing concern over the future directions of Human Factors Engineering in the USAF led to the creation within the Air Force Systems Command (AFSC) of a Human Factors Committee to examine these problems and recommend solutions. One suggestion surfacing through this committee was to transfer all Human Factors psychologists from AFS 2675 to the Medical Specialties corps to enhance the career potential of HFE officers and better control assignments. Other organizations proposed extensive revision of the SEI codes to permit better identification and tracking of subspecializations.

In the context of these various developments, the Air Force Manpower and Personnel Center (AFMPC) classification branch received an unsolicited proposal to conduct an occupational analysis of the behavioral scientist field. This proposal was generated by members of the USAFOMC occupational analysis branch who suggested that it be conducted on a part-time basis, since the total utilization field population did not justify a normal priority occupational survey project. AFMPC approved the request for the project, which had the support of the 2675 Career Development Manager, and an AFPT control number was issued to authorize the survey.

INVENTORY DEVELOPMENT

As a starting point in developing a task list for the 267X field, Air Force personnel documents, such as AFR 36-1, Officer Specialty Descriptions, were screened to identify basic duties and responsibilities of USAF behavioral scientists. In addition, a set of special job descriptions for 267X officers were recovered from AFHRL files. In 1974-75, AFHRL had conducted a special study of all officer AFSCs by collecting narrative job descriptions from a representative sample of positions. Twenty-two behavioral scientist position descriptions were located in this file and served as a foundation for preliminary

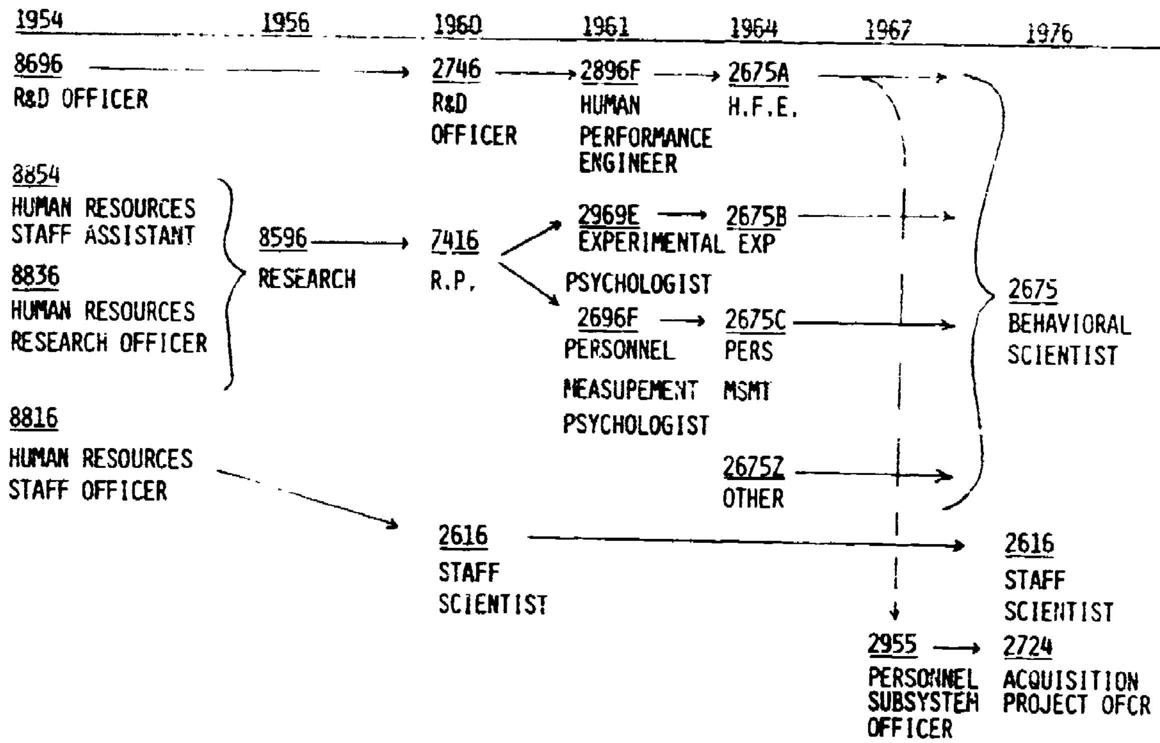


FIGURE 1. EVALUATION OF THE AIR FORCE BEHAVIORAL SCIENTIST UTILIZATION FIELD

AFS 267X USAF JOB INVENTORY DUTY AREAS

DUTY		NUMBER OF TASKS
A	GENERAL COMMAND FUNCTIONS	17
B	SUPERVISORY FUNCTIONS	42
C	ADMINISTRATIVE FUNCTIONS	17
D	GENERAL FUNCTIONS	22
E	PROFESSIONAL DEVELOPMENT FUNCTIONS	14
F	CONSULTANT FUNCTIONS	14
G	LIASION FUNCTIONS	9
H	CONTRACT MONITORING FUNCTIONS	12
I	COUNSELING FUNCTIONS	12
J	RESEARCH FUNCTIONS	22
K	APPLICATIONS OF RESEARCH	20
L	MANAGING RESEARCH OR APPLICATIONS PROGRAMS	14
M	ORGANIZATIONAL IMPROVEMENT FUNCTIONS	11
N	ACADEMIC INSTRUCTOR FUNCTIONS	16
O	HUMAN FACTOR ENGINEERING (HFE) FUNCTIONS	35
P	PROMOTION TEST CONSTRUCTION FUNCTIONS	25
Q	OCCUPATIONAL ANALYSIS FUNCTIONS	28

FIGURE 2. DUTY AREAS FOR THE AFS 267X JOB INVENTORY

task list development. Similar blank forms were reproduced by USAFOMC and mailed to about 30 behavioral scientists to update position descriptions and capture recently developed jobs. As the opportunity presented itself during TDY visits for other purposes (in the normal occupational analysis program), the authors interviewed over 30 behavioral scientists at Wright-Patterson AFB, USAF Academy, Lowry AFB, Keesler AFB, Norton AFB, Gunter AFS, and Randolph AFB. Over one-third of the members of the specialty were contacted either by mail or through personal interviews.

The wide geographical dispersion of jobs and the variety of one and two-deep positions became very obvious during this phase of the project, as did the dynamic nature of the utilization field. New jobs were being created and old jobs deleted quite frequently during this period. A new Lt Colonel's position was created at Gunter AFS to support human factors of computer systems operations. A new 2675 position was created in Air Force Logistics Command headquarters (Wright-Patterson AFB) to evaluate AFLC job enrichment programs. Three or four positions for Captain human factors psychologists were switched to civilian positions at Edwards AFB CA when no qualfied officers (with HFE degrees or experience) could be identified for assignment. The last experimental psychologist position (Armed Forces Radiological Research Institute, Bethesda MD) was deleted upon the incumbent's reassignment. He was replaced by a Veterinarian since no PhD-level experimental psychologist was available. These kinds of shifts in jobs made the development of a comprehensive occupational survey instrument difficult.

Once this phase of the inventory development process was completed, a relatively short task list containing 330 tasks grouped under 17 major duty headings was tentatively developed. Because of the relatively small population (about 140 officer positions) and diverse number of jobs, the task list was written at a more general level of specificity than is normally the case. With small fields such as this, only a few tasks per known job group will adequately differentiate clusters and job types. Thus, a fairly long and detailed task listing was deemed unnecessary.

The Job Inventory was organized functionally, with general duties listed first and more specialized duties following. Duty A, General Command Functions, contained 17 tasks (conduct Commander's Call, etc.) which might be performed by organizational commanders. Duty B listed 42 tasks which most supervisors would perform (Write Officer Efficienty Reports, etc.). Duties C and D outlined Administrative and General Functions which might be performed by most officers, and Duty E detailed Professional Development Functions both for officers in general and for psychologists in particular. The remaining duties (F through Q) related to specific tasks performed by members of some known behavioral science job groups (counselors, researchers, technology applications, HFE, test development, occupational analysis, etc.). These duty headings and the number of tasks for each are displayed in Figure 2.

A fairly extensive background section was also included in the Job Inventory ranging from personal identification, education level, academic specialization, etc. to standard questions concerning job interest and satisfaction (See Figure 3). These types of data facilitate the identification of job groups during analysis and permit a more detailed look at potential problem areas within the utilization field. Such data may be displayed by grade, job type, or organization to highlight differences in groups or to identify particular jobs or areas where morale may be an issue. Finally, the Job Inventory also

2675 BACKGROUND SECTION

I. IDENTIFICATION:

NAME, SSAN, TEL. NO., DOB, SEX, GRADE, MAJOR COMMAND, ORGANIZATION, JOB TITLE, AFIT SCHOOL ATTENDED, AFSCs, TIME-IN JOB, TIME IN AFSC, TAFMS, ENLISTED TIME, COMMISSIONED TIME, NO. SUPERVISED, SOURCE OF COMMISSION, ORGANIZATIONAL LEVEL, SHREDOUT, SEI

II. EDUCATIONAL DATA OF POSITION

REQUIREMENT OF POSITION, SHOULD REQUIRE, ACADEMIC SPECIALTY OF POSITION, ADDITIONAL ACADEMIC CODE

III. JOB REQUIREMENTS

TDY, SUPERVISOR, TYPE OF POSITION, INTERACTIONS WITH VARIOUS CATEGORIES AND LEVELS.

IV. EDUCATION QUALIFICATIONS OF INCUMBENT

DEGREE LEVEL, HOW EARNED, SPECIALIZATION, PROFESSIONAL ORGANIZATION, PME

V. JOB ATTITUDES

CAREER PLANS, ASPIRATIONS, ORGANIZATIONAL CLIMATE, JOB INTEREST AND SATISFACTION

FIGURE 3. BACKGROUND SECTION FOR THE AFS 267X JOB INVENTORY

ORGANIZATIONS REPRESENTED

	NO. SAM
AIR TRAINING COMMAND HQ DCS PLANS (6), TECHNICAL TRAINING HQ & CENTERS (7), AIR UNIV (LMDC/AWC-4), AFIT STAFF & STUDENTS (5), AFROTC (3), USAF RECRUITING SERVICE (1), USAFOMC (41)	6
AIR FORCE SYSTEMS COMMAND AFHRL (12), HQ (3), USAFSAM (1), ESD (1), FTC (3), ASD (3), SAMSO (2), AFAMRL (2), AFWL (1), OTHER (3)	31
USAF ACADEMY	22
DOD AGENCIES (OAR, HR SCHOOL, ETC)	5
OTHER ORGANIZATIONS (TAC, SAC, ETC.)	9

FIGURE 4. ORGANIZATIONS REPRESENTED IN AFS 267X STUDY

included a special section on career assignments in an experimental attempt to determine possible career progression patterns.

The final job inventory was then validated through comprehensive reviews by senior behavioral scientists at AFMPC, AFHRL, and USAFOMC. In addition, the AFMPC Career Development Manager also reviewed the inventory.

FIELD ADMINISTRATION

The printed inventory was administered by direct mailing to organizations which utilize 267X officers, and to those individuals who were in remote locations or one deep positions. Current membership of the utilization field was determined by a roster furnished by the AFMPC Career Development Manager. The inventory was mailed to the field in May and June 1981, with the request for return NLT mid July. In July, follow up telephone calls were made with units and individuals who had not yet responded. Survey administration was closed in August.

Inventory booklets were received from a substantial cross section of the agencies which use USAF behavioral scientists (see Figure 4). Air Training Command has the most behavioral scientists assigned in a variety of diverse jobs; the technical training positions include Technology Applications officers at the five ATC Technical Training Centers. Air Force Systems Command is the second largest user including many one or two-deep Human Factors and research positions. These include the USAF School of Aerospace Medicine, Electronic Systems Division, Flight Test Center, Aeronautical Systems Division, and others. A few positions exist with DOD and the operational commands. In most cases, incumbents held either the 2671, entry level AFSC or 2675 (fully qualified). To insure a complete picture, a small sample of members of other specialties were also asked to participate. These included a small number of Colonels (who cannot hold the 2675) who progressed from behavioral science assignments. Where members of other specialties worked side by side with behavioral scientists, they were also surveyed. This included Education and Training Officers, Instructors, and Staff Scientists. In one case, a Senior Master Sergeant was also included, since he performed the same job as some 2675 officers.

Generally, the sample is very representative of the utilization field, with about 80 percent of all 267X officers included in the study (see Figure 5). A few jobs at some locations are not included for a variety of reasons. Several individuals who were retiring or leaving the USAF declined to participate. In one case at AFHRL at Williams AFB, all three job incumbents were either leaving service or had already departed for reassignment, and replacement officers were not yet in place. While exclusion of these members could lead to some sample bias, the amount of bias is considered to be minimal since 80 percent of the field is captured in the final sample.

PRELIMINARY RESULTS

Since analysis of the survey data has just begun this month, it is not possible to present detailed findings in this paper. However, since there

SAMPLE DISTRIBUTION: GRADE BY DUTY AFSC

AFSC	SMSGT	2LT	1LT	CAPT	MAJ	LTC	COL	TOTAL
2671	0	14	8	13	2	2	0	39
2675	0	2	16	29	22	5	0	74
261X	0	0	0	1	0	2	0	3
0940	0	0	0	2	2	2	0	6
OTHER*	0	1	0	4	1	2	3	11
73200	1	0	0	0	0	0	0	1
TOTAL	1	17	24	49	27	13	3	134

*INCLUDES 0026, 0076, 7516, 7524, ETC.

FIGURE 5. GRADE DISTRIBUTION OF RESPONDENTS

SAMPLE PERCEPTIONS OF ORGANIZATIONAL CLIMATE
(PERCENT RESPONDING)

	0-1/0-2	0-3	0-4	0-5	0-6
DISSATISFIED	45.3	34.7	20.6	35.7	0
NEUTRAL	14.3	0	3.4	0	0
SATISFIED	40.4	65.3	75.9	64.3	100

FIGURE 6. PERCEPTIONS OF ORGANIZATIONAL CLIMATE

are several high interest items, such as job satisfaction and career plans, which can be easily addressed at this time, some preliminary results are presented for these areas. In addition, a brief overview of the job structure will be discussed.

Figure 6 displays the perceptions of behavioral scientists in various grades of the organizational climate of their unit. As you will note, there appear to be some substantial differences among the grade groups, with Lts being generally less satisfied than other groups. This is perhaps not a startling conclusion; however, we also see basically the same pattern when we examine responses to the question on job interest (see Fig. 7). Again, the extreme responses seem to be in the Lts and Colonel groups. A slightly different pattern was noted for the question as to how the job utilizes ones talents (see Fig. 8). Here, Majors seem to be most satisfied, with the Lts again having a greater proportion of dissatisfied individuals.

When we asked what sense of accomplishment behavioral scientists received from their jobs, the pattern was similar to earlier questions (see Fig. 9). Again, this is hardly a surprising finding, since the Lts group contains some individuals who may not choose a full military career. We hope to be able to develop a comparative data base from another officer study, the Professional Military Education project (Barucky 1980), to compare behavioral scientists with the Air Force population at large. Those results will be presented at a later time.

Figure 10 displays career plans by major organizational unit. As you can see from this data, a majority of the individuals in Air University and the USAF Academy plan to remain in behavioral science careers, while a substantially lower percentage of officers assigned to Air Training Command have similar plans. This may be a function of the larger population in ATC (N=67), with a higher proportion of very junior officers. This is the kind of question we will be examining in detail as we get further into the analysis of these data.

One of the major objectives of this study was to identify the major types of jobs which exist in the utilization field. We were able to use a new experimental CODAP routine called CORSET, which Mr. Bill Phalen and Mr. Johnny Weissmuller are reporting in another session of this conference (Phalen & Weismuller 1981). This program permits a very quick analysis of the tasks which discriminate the various groups formed in the hierarachical clustering process to establish which are significant groups. Using this new technique, we were able to collapse the 38 starter groups on our diagram (W=2) into 25 job types which clustered into seven major clusters. These include: Research Programs Officers (35% of the sample); Functional Unit Supervisors (10%); Academic Instructors-Counselors (10%); Junior Task Scientists and Students (9%); Occupational Analysts (15%); Human Factors Engineering Researchers (7%); and Test Development Psychologists (12%). These major types of jobs include 98% of the sample. The remaining individuals were filling a number of one-deep, unique positions.

Each of the major clusters is composed of more discrete job types, which represent variations of tasks. For example, the Research Program Officers cluster includes the following groups: Personnel Research Program Managers, Technology Application Researchers, Technology Applications Staff Officers, Plans Staff Officers, Senior Academic Staff Officers, Contract Monitors, Test Development Research, MPC Attitude Researchers, and Air War College Evaluation research. As

HOW DO YOU FIND YOUR JOB?

(PERCENT RESPONDING)

	LTS	CAPT	MAJ	LTC	COL
DULL	21.4	8.1	20.6	0	0
SO-SO	7.1	4.1	0	7.1	0
INTERESTING	71.5	87.8	79.4	92.9	100

FIGURE 7. PERCEPTIONS OF JOB INTEREST

HOW DOES YOUR JOB UTILIZE YOUR TALENTS?

(PERCENT RESPONDING)

	LTS	CAPT	MAJ	LTC	COL
NOT AT ALL OR VERY LITTLE	31.0	16.3	10.3	0	0
FAIRLY WELL TO VERY WELL	57.1	51.0	41.3	57.1	66.7
EXCELLENTLY TO PERFECTLY	11.9	32.7	48.3	42.9	33.3

FIGURE 8. PERCEPTIONS OF JOB USE OF TALENTS

SENSE OF ACCOMPLISHMENT FROM YOUR JOB?

(PERCENT RESPONDING)

	LTS	CAPT	MAJ	LTC	COL
DISSATISFIED	26.1	14.3	13.7	7.1	0
NEUTRAL	2.4	2.0	6.9	7.1	0
SATISFIED	71.5	83.7	79.3	85.7	100

FIGURE 9. PERCEPTIONS OF ACCOMPLISHMENT FROM JOB

CAREER PLANS?

(PERCENT RESPONDING)

	USAFA	AFSC	ATC	AU	HQ USAF
STAY IN 2675	54.5	41.9	28.4	75.0	0
STAY IN 26XX	18.2	12.9	19.4	0	33.3
CROSS TRAIN	0	19.4	19.4	0	0
UNDECIDED	4.5	16.1	17.9	0	0
SEPARATE BEFORE RETIREMENT	4.5	0	4.5	0	0
RATED SUPPLEMENT - CAREER BROADENING	18.2	6.5	3.0	0	33.3
OTHER PLANS	0	6.5	4.5	25.0	33.3

FIGURE 10. SUMMARY OF CAREER PLANS OF AFS 267X OFFICERS

might be surmised from the various titles, each group appears to specialize in research activities involving a different organizational mission or program and perform a slightly different set of specific tasks.

The Academic Instructor-Counselor cluster is composed of the following groups: USAF Academy Instructors; USAFA Instructor-Counselors; AFROTC Instructor-Counselors; and Other Instructors. As these names imply, some USAFA faculty members perform primarily as classroom teachers (another group identified in the first major cluster perform both classroom instruction and conduct research projects). A separate group both instructs and serves as counselors. This group includes individuals assigned to the Cadet Counseling Center, which is now part of the Department of Behavioral Science and Leadership. Their "core" tasks suggest that they are being used both as counselors and as classroom instructors. Interestingly, the AFROTC instructors included in this study are more similar to the USAFA Instructor-Counselors than to the Instructor group. Their "core" tasks reflect considerable personal counseling outside of the classroom. The Other Instructor group includes several unique, one-of-a-kind faculty positions, such as the DOD Human Relations Institute at Patrick AFB, FL. The group of other instructors seem to group together by their shared instructing, counseling, and course development tasks.

Most of the other major job groups involved highly specialized functional programs. In the Occupational Analysts group, there were three distinct job types which included Airman Analysis (OMYO), Inventory Development (OMYV), and Management Applications and Officer Analysis (OMYA) personnel, thus replicating the formal sections of our branch (some section chiefs grouped in the supervisor job group). Interestingly, the present USAF-Royal Australian Air Force Exchange officer grouped with OMYA analysts, even though the incumbent has never been assigned to the USAFOMC. This is very realistic since we know that he both does inventory development and data analysis as do the members of OMYA. One member of the Canadian Armed Forces occupational analysis program also appeared in the job group (he completed the survey in place of the newly assigned USAF-Canadian Forces exchange officer who reported for duty in July). These groupings were particularly satisfying since they give some external validity to our survey results.

The Human Factors Engineering Researchers formed a very discrete job group with very little overlap with other groups (although their supervisors and staff personnel did appear in the first cluster). Members of this group represent very different organizations including ASD, ESD, the Flight Test Center, and others. Their "core" tasks reflect a concentration of 6.3 and 6.4 development efforts not shared by any other group.

Finally, the WAPS Test Development Psychologists assigned to the USAFOMC have very distinct jobs. Two job types within this cluster include test development psyhcologists and test review psychologists. Their jobs focus on procedures for developing promotion Specialty Knowledge Tests and the quality control of such tests.

It is not possible in this brief, preliminary report to examine these job groups in any great detail. A more complete report of the occupational analysis project is to be published next spring, and we hope to be able to give a more detailed report at a later date (perhaps the Psychology in the DOD Symposium at the USAF Academy in April, or some other forum).

It is possible, however, to conclude from the preliminary results of this occupational analysis that the behavioral scientist is a very diverse utilization field. A quick look at job perceptions of incumbents suggests that there may be some problems, particularly among the more junior officers, and that these problems may vary by major command or organization. We are hoping that this occupational analysis may help solve some of the problems through better identification of jobs and highlighting potential areas requiring action. Before such results can be achieved, a considerable amount of further analysis is required.

References

Barucky, J.M., Officer Professional Military Education Curriculum Validation Project. Occupational Survey Report, Occupational Analysis Program, USAFOMC, Randolph AFB, TX, 1980.

Christal, R.E., The United States Air Force Occupational Research Project. AFHRL-TR-73-75, AD-774 574. Lackland AFB, TX: Occupational Research Division, Air Force Human Resources Laboratory, January 1974.

Jacoby, Jacob., "The Plight of the Uniformed Air Force Psychologist." Professional Psychology, 1970, SUMMER, Vol 1(4), 383-387.

Mitchell, J.L. and Driskill, W.E., Variance Within Occupational Fields: Jobs Analysis Versus Occupational Analysis. Paper presented at the 21st Annual Conference of the Military Testing Association, San Diego, CA, 1979.

Morsh, J.E., "Job Analysis in the United States Air Force." Personnel Psychology, 1964, 17(1), 7-17.

Phalen, W.J. and Weissmuller, J.J., CODAP: Some New Techniques to Improve Job-Type Identification and Definition. Paper presented at the 23rd Annual Conference of the Military Testing Association, Washington, D.C., 1981.

AD P001356

Modrick, John A., Plocher, T. A. & Hutcheson, J. D., Honeywell, Inc., Minneapolis, Minnesota; Chambers, R. M., US Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia. (Tues. A.M.)

Performance and Skill Level Requirements for Fire Support Teams

This research developed a Task Data Base for Fire Support Teams (FIST) and identified 48 tasks, grouped them into six functional areas, classified them as either procedural or semi-structured, and ranked them according to criteria of criticality and performance. The FIST Task Data Inventory resulted from the task analyses and integration of task descriptions and information obtained from questionnaires, interviews, and observations administered at three CONUS and four USAREUR divisions. Statistical analyses of the data indicated that (1) criticality and performance were negatively correlated, (2) procedural tasks were the least critical and best performed, (3) semi-structured tasks were the most critical and poorest performed, (4) task difficulty was the principal factor in ratings of criticality, and (5) traditional tasks were performed better than non-traditional tasks. The results of these analyses, and the utilization of the FIST Task Data Inventory, are discussed in terms of personnel and training assessments, simulation and training device recommendations, task analyses methodologies, and selection criteria.

PREVIOUS PAGE IS BLANK

PERFORMANCE AND SKILL REQUIREMENTS FOR FIRE SUPPORT TEAMS

J.A. Modrick, T.A. Plocher, J.D. Hutcheson

Systems and Research Center
Honeywell Inc.
Minneapolis, Minnesota

R.M. Chambers

Army Research Institute
Alexandria, Virginia

The purpose of this research was to examine the needs and problems in manpower and training which have arisen during implementation of the Fire Support Team (FIST). The work was performed under Contract No. MDA903-99-C-9669 with the Army Research Institute, November 1, 1979, to March 30, 1981.

The FIST is a new entity which replaces the traditional forward observer by a team which is intended to provide wider coverage, greater mobility, and effective integration of improved munitions. The FIST consists of forward observer parties deployed with platoons to provide flexibility, mobility, and range, and a headquarters element at the company command post to provide integration, coordination, planning, and responsibility for the company's scheme of maneuver.

FISTs have been formed by training new personnel, by reclassification, and by assignment to the field. This process required reorganization and new development of training at both resident school and unit.

OBJECTIVES AND APPROACH

There were three objectives of study: 1) Specify the performance and skill level requirements for personnel assigned to FISTs; 2) Determine the degree to which performance and skill level requirements are being met, and identify the shortfalls which exist in manning, organization, equipment, and training of FISTs in the field; and 3) Project the probable impact that future field artillery systems will have on the performance and skill requirements for FIST.

The approach consisted of five tasks: 1) Develop a task data base for FIST; 2) Collect interview and questionnaire data from fire support personnel in fielded units in CONUS and USAREUR; 3) Analyze interview and questionnaire data for problems or deficiencies in performance, and training needs and contributing factors; 4) Estimate the effect of future weapon systems on job task and training requirements for FIST; and 5) Summarize FIST training and performance deficiencies, and make recommendations to correct them.

DEVELOPMENT OF A TASK DATA BASE FOR FIST

A Task Data Base was compiled by integrating several sources; no single source provided an adequate list of tasks or adequate task data. It was necessary also to supplement this information by interviews with experts and by analysis. The printed sources of task-descriptive information were: Soldiers' Manuals FM6-13F 1/2 and 3; ARTEP 6-365 FA Battalion, 155mm DS; ARTEP 6-105 Battalion, 105mm DS; Training Circular 6-20-10 FIST, Task Analysis 13F, Directorate of Course Development DCRDT; Task Analysis, Close-Support Study Group II; Training Text 6-20-7, FIST/FAC Operations; and Weapon S stems Training Effectiveness Analysis - Forward Observer. The resulting list of tasks provided a comprehensive inventory of job tasks. It constituted a data base which was used to structure interviews, performance analyses, and training analyses in the later phases of the study.

The data on each task was summarized on a task description worksheet which contained the following categories of information: task number; duty (major functional area) to which the task belongs; task description; task assignment to positions within FIST, as indicated or implied by doctrine and as implemented in current field practice; task criticality rank and narrative evaluation of task criticality; references for task information; and listing of component subtasks, if applicable, and references to the source of subtask descriptions. These data were obtained by analysis of responses to a questionnaire and interviews. Functional flow diagrams were also prepared for nonprocedural, semi-structural tasks, or a set of related tasks, which were not amenable to description by a fixed sequence of subtasks or steps.

COLLECTION OF INTERVIEW AND QUESTIONNAIRE DATA

Questionnaires were administered and interviews conducted among personnel of FIST units in the field, related fire support organizations, and commanders of maneuver units. The purpose of the questionnaire was to obtain information on the military occupational specialty (MOS) background of FIST personnel, nature of a company commander's experience with the FIST, and ratings of tasks on criticality and quality of performance by the FIST. The purpose of the interview was to obtain less easily structured information such as factors and considerations influencing a respondent's ratings and perceptions of the adequacy of training.

The questionnaire consisted of Background and Task Inventory sections. Different forms were used for FIST personnel and the company commander. For FIST participants it consisted of length of service, MOS, and military training history; for company commanders it consisted of the nature of their exposure to FIST concept prior to assuming a company command. The Task Inventory section was designed to obtain evaluations of each FIST task on five attributes: Task Assignment, in terms of the FIST member responsible for the task; Performance Rating from combat-ready to totally inadequate; and Tast Criticality based on three estimates. They were: Task Difficulty (not difficult to extremely difficult); Consequences of Inadequate Performance (catastrophic to none); and Detectability of Error (undetectable to immediately detectable by the person committing the error). Criticality was not intended to reflect importance of a task but properties of tasks that are significant in determining priorities for training.

Only FIST Chiefs made evaluations on all five scales. FIST enlisted personnel responded to Task Assignment with a Yes or No to indicate whether they had performed that task in their assigned job and rated task difficulty; company commanders rated only performance and consequences of inadequate performance.

Interviews were conducted in three divisions in the continental U.S. and three in Europe. Questionnaires were administered to 86 and 137 persons in the U.S. and Europe, respectively; interviews were conducted with 67 and 94 persons. An interview protocol was prepared for each category of respondent to be used as a checklist of coverage to guide the interview. A primary concern was to allow the respondent latitude in his responses and comments and permit the interviewer freedom to follow up and elaborate on ratings and comments.

The results are organized into Task Data, Training, Personnel, and Organizational Factors.

FIST TASK DATA BASE

The information from the printed sources was combined with task data obtained by analysis of responses to the questionnaires and interviews. Forty-eight tasks were identified and grouped into six functional areas or duties. The functional areas and number of tasks in each are: I. Plan fires to support maneuver unit operations (9); II. Prepare/maintain/disseminate fire support information (6); III. Manage fire support communication system (5); IV. Manage use of fire support assets at maneuver unit level (3); V. Acquire targets of opportunity (4); and VI. Request/adjust fires (21). Information flow diagrams were prepared for some tasks and sets of tasks to depict the interdependencies among the tasks. They can be used as a basis to plan and operate simulations and training exercises for the purpose of integrating the component activities into an effective operation.

A set of basic skills too molecular to be components of tasks emerged as tasks were analyzed in greater detail. They are basic skills of fire support and are common prerequisites to mastering several tasks representing entry-level skills. The skills were grouped into the following six categories: Basic Forward Observer Procedures; Intelligence Reporting; Basic Map Reading; Basic Radio Procedures; Maintenance; and Basic Fire Support Procedures.

The tasks were also classified as procedural and semi-structured. Procedural tasks have a fixed sequence of steps or operations, initiating conditions and contingencies are known, standards of performance are known in the form of one correct response or outcome. They occur in established situations. Semi-structured tasks have a sequence of steps or operations which varies depending on outcome of prior step, initiating events and conditions vary with situations, standards of performance are not known or determined, and multiplicity o appropriate responses exists. They occur in emergent situations. Current training technology is oriented predominantly toward procedural tasks while semi-structured tasks are new to the fire support community. They involve planning, coordination, and integration of fire support.

Mean performance and criticality ratings were computed for each task and the tasks were put in rank order. The ratings of the three criticality scales were combined to provide a single composite index of criticality. The relationship between the rank order of performance and criticality was estimated by Spearman Rho and found to be -0.67. Perceived level of performance declines as the tasks increase in criticality.

The tasks were divided into thirds on each dimension. The most critical and poorest performance (MC/PP) third and the least critical and best performed (LC/BP) third of the tasks were analyzed. The LC/BP tasks are procedural with one exception. The MC/PP tasks are semi-structured and procedural tasks on which FIST personnel get little practice or experience such as Request/Direct Close Air Support.

Task difficulty is the major factor of the criticality ratings. The partial correlation between performance ratings and task difficulty is 0.74 ($p < 0.001$) controlling for the covariance with consequences of inadequate performance and detectability of errors. Comparable partial correlations of performance ratings with consequences and detectability of errors are 0.03 and -0.33 ($p < 0.05$) respectively. The tasks for which errors are more detectable are performed better.

The correlation between the criticality dimensions indicate independence among these dimensions. They are: task difficulty/detectability (-0.22); task difficulty/consequences of inadequate performance (0.30, $p < 0.05$); and detectability of errors/consequences of inadequate performance (-0.33).

TRAINING

Resident school training is viewed as good and unit training as poor. Training courses provided at the Field Artillery School in Officer's Basic, Cannon Battery Officers', Basic NCO, and Advanced Individual Training are good. However, they focus largely on the procedural tasks; less proceduralized tasks in planning, coordinating, and integrating fire support are not adequately covered or exercised. Unit training programs are not adequate in FIST participation in field exercises for maneuver units, unit training for FIST, and reclassification training.

The characterization of unit training provided by the survey is represented in the following responses. Seventy-three percent of enlisted respondents in CONUS and USAREUR reported that MOS unit training is not sufficient to become proficient at one's job and skill level. Over half the respondents cited training shortfalls as a major source of performance deficiency. An average of 9-2 hours/week of scheduled training activity while in garrison was reported by 60% of USAREUR respondents; CONUS observations were similar. Deviations took the form of "crash courses" prior to SQT and ARTEP times. Seventy percent of CONUS/USAREUR reclassified respondents reported no transition in unit other than OJT.

The training methods reported as currently used to train in support planning consist of: the FTX (Field Training Exercise), consisting of ARTEP, REALTRAIN, and GDP Areas Walks; the CPX (Command Post Exercise) consisting of CAMMS, Battle of Eiterfeld, and Dunn Kempf; and classroom lectures. The typical USAREUR annual FTX training consists of supporting one battalion ARTEP, one three-week REALTRAIN exercise at Hohenfel MTA, and one/two GDP terrain walks. Factors limiting the value of the FTX in USAREUR are the severe constraints on field maneuver, physical separation of units up to 100 KM, no feedback on effect of fire support planning, unrealistic representation of time constraints and availability of assets and nonrepresentation of multiple information sources.

The value of the CPX for FIST is compromised by limitations of the simulation. Fire support planning is infrequently exercised. CAMMS is the most commonly used but the company level and FIST are not exercised in CAMMS. The Dunn Kempf game, a company level simulation, provides training for FIST HQ and the FO team, but with mixed results. It has questionable validity of weapons effects and requires excessive time to learn rules and to prepare/set up/run. The Battle of Eiterfeld designed in December 1979 to meet FIST needs has not yet reached the field units.

One good program exists in 8th Infantry DIVARTY. It is a three-phased program progressing from classroom instruction through CPXs to a combined arms exercise in live firing. The phases are: 1) Four concentrated hours of class on fundamentals of fire support planning; 2) Three series of Dunn Kempf gaming exercises played; and 3) Combined arms live fire exercise and evaluation.

The principal conclusions concerning unit training is that the FIST needs a separate exercise capability designed to meet its training needs which can be used to provide frequent training experiences. It also needs the resources and autonomy to manage its own training and development of its personnel.

PERSONNEL

A summary of FIST personnel factors is: FISTS in the field are understaffed or staffed with underqualified personnel in numbers, grade levels, training, and experience; there is a potential problem of retention; and personnel authorizations may not be adequate in experience and depth to sustain prolonged performance.

Manpower shortfalls exist in both enlisted and officer ranks. Enlisted ranks are operating at 65% of authorized strength ranging from 38% to 86%. The E7 grade showed the most marked shortfall. Officer ranks are operating at 60% of authorized strength, ranging from 46% to 93%. The Brigade FSO position is rarely filled by an O4 grade; the O3 grade was only slightly more common in the Battalion FSO positions. The pool of FIST lieutenants was significantly understrength. FIST personnel are routinely forced to perform in job positions intended for higher grade/rank and more experience. FISTs are routinely sent to the field with less than a complete crew in number, MOS qualification, and grade/rank.

There is a credibility problem for the FIST in many units. Many company commanders do not value FIST or know how to use it, and do not treat it as part of the command team. A similar gap exists at the level of FO team and platoon leader. Since deployed FISTs are only a partial implementation of the doctrinal concept and company commanders routinely see reclassified E5s/E6s as FIST chiefs. limited use of the FIST chief as fire support coordinator is a safer course.

Most reclassified FISTs come from MOS related primarily to forward observer positions. Few reclassified FIST members have had reclassification training. Thirty-seven percent of the E1-E5 personnel reported lack of expertise among 13F NCOs as a major impediment to training.

Common sources of dissatisfaction in MOS 13F are that they routinely work in a "shortage" environment and are required to perform in jobs for which they have not qualified. Further, capable, undergrade FIST personnel often express frustration at being underutilized and unappreciated. The field artillery battalion is still focused on guns and the fire direction center. Career and training needs of the 13F specialists are often subordinated to priorities of firing batteries.

ORGANIZATIONAL FACTORS

There are needs for greater self-sufficiency for FIST in managing its own training, resources, and personnel, and to emphasize a forward-looking rather than reactive role in the management and integration function of FIST. There is doubt that the present FIST implementation provides sufficient resilience and flexibility for operation under emergency or degraded conditions.

Eighty percent of FIST officers and enlisted respondents in USAREUR and 90% in CONUS expressed dissatisfaction with current FIST utilization. Perceived misuse of the fire support section within battalion often was cited as a major impediment to training. The fundamental problems faced by fielded units are effective integration into the battalion organization, lack of an MOS-experienced 13F NCO pool, and lack of time due to excessive support details. Eighty percent of FIST reported this misuse of the FIST as a serious factor. Three alternatives proposed to organizing the FIST were: attach all

FIST personnel to the firing battery; attach some or all FIST lieutenants to the firing batteries with FIST enlisted personnel remaining in the HQ battery; and consolidate all FIST officers and enlisted personnel in HQ battery under the Brigade FSO.

IMPACT OF FUTURE SYSTEMS ON THE FIST SUPPORT TEAM

A summary of the implications of future systems for FIST training must emphasize the following points. There will be an increased complexity of HQ tasks in fire support planning, management of fire support resources, and support to scheme of maneuver. They are the semi-structured tasks which are the less well trained tasks at the present time. They are the more difficult to train and require "hands-on" exercises as a necessary method. Training will require greater use of simulation, modeling, and war gaming in training devices and programs. These methods are costly and not mature technologically. Future systems will have a lesser effect on the functions of the forward observer.

Recommendations for research and development were made in three areas: Need for more knowledge and better understanding of FIST as a team, its performance, and the workload imposed by various combat scenarios; need for improved training materials and delivery systems; and improved retention in the 13F MOS.